Safe Exploration via Policy Priors

Manuel Wendl, Yarden As*
ETH Zurich

Manish Prajapat
ETH Zurich

Anton Pollak
ETH Zurich

Stelian Coros
ETH Zurich

Andreas Krause
ETH Zurich

Abstract

Safe exploration is a key requirement for reinforcement learning agents to learn and adapt online, beyond controlled (e.g. simulated) environments. In this work, we tackle this challenge by utilizing suboptimal yet conservative policies (e.g., obtained from offline data or simulators) as priors. Our approach, SOOPER, uses probabilistic dynamics models to optimistically explore, yet pessimistically fall back to the conservative policy prior if needed. We prove that SOOPER guarantees safety throughout learning, and establish convergence to an optimal policy by bounding its cumulative regret. Extensive experiments on key safe RL benchmarks and real-world hardware demonstrate that SOOPER is scalable, outperforms the state-of-the-art and validate our theoretical guarantees in practice.

1 Introduction

A defining feature of intelligence is the ability to utilize online streams of information to learn and adapt over time [1]. Reinforcement learning (RL) provides a framework for online learning without supervision, driving several notable real-world applications [2, 3]. Safety is key to unlocking RL in the physical world—RL agents simply cannot risk taking actions that lead to catastrophic outcomes [4], even *while* learning, when they lack complete knowledge about their environments [5]. Such agents must *explore safely* [6], gradually discovering new safe behaviors to solve their tasks.

Safe exploration is a major challenge in practice. Methods with theoretical safety guarantees often struggle to scale to complex, general-purpose tasks, whereas scalable approaches typically fail to ensure safety during learning (cf. Section 2). A key reason lies in how they incorporate prior knowledge about the environment—a crucial component for anticipating danger without learning through harmful trial-and-error. Such prior knowledge can be instantiated via *policies*. Beyond primarily maintaining safety, as done in much of prior work, we argue that prior policies can also *guide* exploration toward promising regions of the environment, enabling agents to provably learn *near-optimal* policies.

Motivated by this insight, we propose **S**afe **O**nline **O**ptimism for **P**essimistic **E**xpansion in **R**L—SOOPER—a model-based RL algorithm for safe exploration in constrained Markov decision processes [CMDP, 7]. SOOPER uses prior policies that can be derived from scarce offline data or simulation under distribution shifts. Such policies are invoked *pessimistically* during online rollouts to maintain safety. These rollouts are collected *optimistically* so as to maximize information about a world model of the environment. Using this model, SOOPER constructs a simulated RL environment, used for planning and exploration [6], whose trajectories *terminate* once the prior policy is invoked. The key idea is that early terminations incentivize the agent to avoid the prior policy when trajectories with higher returns can be obtained safely. This design allows SOOPER to leverage standard RL methods, bypassing the complexity of directly solving CMDPs. More fundamentally, this enables a novel bound on the agent's cumulative regret.

[*] Equal contribution. Corresponding authors: {mwendl,yardas}@ethz.ch

2nd Workshop on Aligning Reinforcement Learning Experimentalists and Theorists (ARLET 2025).

Our contribution.

- First, we propose SOOPER, a scalable model-based RL algorithm for safe exploration in continuous state-action spaces that leverages offline- and simulation-trained policies as safe priors.
- Next, under standard regularity assumptions, we prove that SOOPER guarantees constraint satisfaction with high probability throughout learning. We additionally show that SOOPER gradually expands an implicit set of safe policies until it converges to the feasible set defined by the true CMDP. This is accomplished by reformulating the problem as an unconstrained MDP, making the approach naturally compatible with standard deep RL methods, while avoiding min-max formulations typical for CMDPs. Using this formulation, we further establish a novel upper bound on the cumulative regret, improving over prior works that provide optimality guarantees only at the end of training with arbitrarily poor performance during exploration [8, 9].
- Finally, we conduct extensive empirical evaluation of SOOPER, comparing it with other state-of-the-art safe exploration algorithms when initialized from offline- and simulation-trained policies in RWRL benchmark [10] and SafetyGym [11]. In all these experiments, SOOPER substantially outperforms the baselines in terms of safety and performance. We further validate SOOPER for safe online learning on *hardware*, giving empirical evidence that our theory translates to the real world.

2 Related Works

Prior works on safe exploration fall broadly into two categories: theoretically principled methods with guarantees on safety and optimality and those prioritizing practicality and scalability.

Provable safety. A canonical line of work leverages safe Bayesian optimization [12], which has seen real-world success [13, 14, 15], but typically ignores the Markovian structure of sequential decision problems and is limited to a few parameters. Berkenkamp et al. [16] extend to continuous MDPs and provide safety and optimality guarantees via Lyapunov stability, though they rely on manual resets of the system at each timestep and state-space discretization for safe set computation. In contrast, Prajapat et al. [17] leverage model predictive control (MPC) to implicitly compute these sets and ensure that the agent can always return to a safe state. Other MPC-based methods ensure safety either through "backup" policies [18, 19, 20] or via low-fidelity models with reachability analysis [21] and safety filters [22, 23]. However, these approaches lack optimality guarantees. A substantial body of work addresses safe exploration in *discrete* CMDPs [24, 25, 26, 27, 28, 29, 30]. Akin to our work, Bura et al. [30] proposes a model-based approach that uses optimism and pessimism to provide optimality and safety guarantees for small-scale tabular problems. Although the above methods guarantee safety throughout learning, scaling them to high-dimensional problems is computationally hard.

Scalability. The works of Dalal et al. [31], Eysenbach et al. [32], Srinivasan et al. [33], Thananjeyan et al. [34], Bharadhwaj et al. [35], Thomas et al. [36], Sun et al. [37], Liu et al. [38], As et al. [39], Sootla et al. [40], Huang et al. [41] use scalable tools from deep RL for safe learning, with different levels of theoretical guarantees and effectiveness in maintaining safety throughout learning in practice. Safe exploration has also been studied from an *optimization* perspective, with methods aiming to guarantee feasibility of all optimization iterates. These techniques span trust regions [42, 43], Lyapunov stability [44], interior-point [45, 46, 47], and primal-dual optimization [48]. While these methods scale well, they often overlook the exploration aspect of learning, namely how to actively sample data that maximizes information about the environment. As a result, they often lack optimality except under assumptions on the initial state distribution.

Optimality. As et al. [9] propose ActSafe, which guarantees safety and optimality for *simple* regret, relying on reward-free exploration, and therefore might perform poorly during exploration. Wachi et al. [49] introduce MASE, which enforces safety under high-probability constraints on safe states but relies on a restrictive "emergency stop" action that can effectively 'stop time' at any state. Lastly, Wagener et al. [8] propose SAILR, a state-of-the-art algorithm that like SOOPER, reformulates safety through terminations whenever their "backup" policy is invoked. However, its guarantee on simple regret hinges on the probability of such resets vanishing over time, yet no formal conditions are provided to ensure this occurs. In this work, we relax several of these assumptions and establish a novel bound on the cumulative regret, while also demonstrating improved empirical performance.

3 Problem Setting

Constrained Markov decision processes. Safety in RL has been conceptualized in several ways across the literature [50, 51, 52]. In this work, we focus on constrained Markov decision processes [CMDP, 7], due to their flexibility and capacity to generalize several notable formulations of safety [cf. 8, 23, 49]. A discounted infinite-horizon CMDP is defined by the tuple $\mathcal{M}_c := (\mathcal{S}, \mathcal{A}, p, r, c, \gamma, \rho_0)$, where $\mathcal{S} \subseteq \mathbb{R}^{d_\mathcal{S}}$ and $\mathcal{A} \subset \mathbb{R}^{d_\mathcal{A}}$ are the continuous state and compact action spaces, with the transition probability $p(s_{t+1}|s_t, a_t)$ of reaching $s_{t+1} \in \mathcal{S}$ by applying $a_t \in \mathcal{A}$ in $s_t \in \mathcal{S}$. We consider the transition probabilities p to be governed by the *unknown* stochastic dynamics

$$s_{t+1} = f(s_t, a_t) + \omega_t, \quad (s_t, a_t) \in \mathcal{S} \times \mathcal{A}, \quad s_0 \sim \rho_0, \tag{1}$$

with the initial state distribution ρ_0 and additive noise $\omega_t \in \mathcal{W} \subseteq \mathbb{R}^{d_\mathcal{S}}$. Throughout our theoretical analysis, we consider the setting of additive zero-mean Gaussian noise.

Assumption 1 (Gaussian noise [53])**.** *The additive noise ω_t is independent and identically distributed (i.i.d.) zero-mean Gaussian noise with variance σ^2.*

We note that this assumption can be relaxed to a richer class of sub-Gaussian distributions. The reward and the cost are given by $r : \mathcal{S} \times \mathcal{A} \to [0, R_{\text{max}}]$ and $c : \mathcal{S} \times \mathcal{A} \to [0, C_{\text{max}}]$ with discount factor $\gamma \in (0, 1)$. We assume the following regularity conditions for the dynamics, reward, and cost.

Assumption 2 (Lipschitz-continuity [16])**.** *The dynamics f, reward r and cost c are Lipschitz continuous with known constants L_f, L_r, L_c respectively.*

Assumption 3 (Regularity [16])**.** *The unknown dynamics f lie in an RKHS associated with a kernel k and have a component-wise bounded norm $||f_i||_k \leq B$ with a known constant B. Moreover, we assume the kernel k to be bounded by $k((s, a), (s, a)) \leq k_{max}, \forall (s, a) \in \mathcal{S} \times \mathcal{A}$.*

Intuitively, this assumption means that the complexity of f is controlled by the kernel, which in turn enables generalization from limited data. Our analysis uses Assumption 3 to relate the statistical complexity of learning the dynamics to the number of online episodes required for convergence. In Section 5, we model it using neural networks that can learn features to represent f accurately.

The goal in our setting is to find a stationary, stochastic policy $\pi_c^*(a_t|s_t)$ for the true dynamics f that maximizes value function $V_r^\pi(s) := \mathbb{E}_\pi[\sum_{t=0}^\infty \gamma^t r(s_t, a_t) \mid s_0 = s]$, while ensuring that the cost value function $V_c^\pi(s) := \mathbb{E}_\pi[\sum_{t=0}^\infty \gamma^t c(s_t, a_t) \mid s_0 = s]$ remains below a safety budget $d \in \mathbb{R}_{\geq 0}$:

$$\pi_c^* \in \arg\max_\pi J_r(\pi, f) := \mathbb{E}_{s_0 \sim \rho_0}\left[V_r^\pi(s_0)\right] \quad \text{s.t.} \quad J_c(\pi, f) := \mathbb{E}_{s_0 \sim \rho_0}\left[V_c^\pi(s_0)\right] \leq d. \tag{2}$$

We hereafter consider in our theoretical analysis strict feasibility of π_c^*; namely that an optimal policy of the constrained problem does not lie on the boundary of the feasible set but may be arbitrarily close, such that $J_c(\pi_c^*, f) < d$. Equation (2) naturally decouples the objective from safety, so each can be solved independently. This explicit separation prevents agents from exploiting the reward function to achieve high performance at the expense of harmful behaviors, a problem commonly referred to as *reward hacking* in the AI safety literature [6].

Safe online learning. Since knowledge of the dynamics is limited a priori, the agent must learn them through online interaction. We consider learning in an *episodic* setting, where in each episode $n = 1, \ldots N$, infinite-horizon trajectories are truncated after T_n steps and the agent is reset to a new initial state [54, 55]. The agent deploys a policy π_n, collecting transitions $\mathcal{D}_n = \{(s_t, a_t, s_{t+1})_{t=1,\ldots T_n}\}$. The optimality gap due to incomplete knowledge over episodes is quantified by cumulative regret, namely, the difference in performance between π_n and the optimal policy given full access to $\mathcal{M}_c$:

$$R(N) = \sum_{n=1}^N \left(J_r(\pi_c^*, f) - J_r(\pi_n, f)\right) \quad \text{s.t.} \quad J_c(\pi_n, f) \leq d, \, \forall n \in \{1, \ldots, N\}. \tag{3}$$

The crux of safe exploration is that unlike performance, where suboptimality during learning is tolerable, safety must be maintained throughout all n episodes.

Probabilistic world models. Given full knowledge of the CMDP $\mathcal{M}_c$, the agent can solve Equation (3) with no regret. In practice however, the agent has only limited knowledge about $\mathcal{M}_c$. We capture this uncertainty by defining a *set of plausible models* of the dynamics $\mathcal{F}_n$ for each iteration n.

Definition 1. *We define the set of plausible models in episode n' as $\mathcal{F}_{n'} := \{\tilde{f} \mid |\tilde{f} - \mu_{n'}| \leq \beta_{n'}\sigma_{n'}\}$, described by a nominal model $\mu_{n'} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^{d_{\mathcal{S}}}$ and the uncertainty $\sigma_{n'} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^{d_{\mathcal{S}}}$ given data $\mathcal{D}_{1:n'}$. This implies the model of f is always well-calibrated [56], if there exist $\forall n' \leq n : \beta_{n'} \in \mathbb{R}_{>0}$ such that, with probability of at least $1 - \delta$ it holds $\forall (s, a) \in \mathcal{S} \times \mathcal{A}$ that $f(s, a) \in \mathcal{F}_n := \bigcap_{n'=0}^{n} \mathcal{F}_{n'}$.*

Definition 1 formalizes the agent's *epistemic* uncertainty about $\mathcal{M}_c$. At each iteration n, the agent maintains a confidence set that is statistically consistent with the observed data. This is done by learning a model with nominal dynamics $\mu_n : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^{d_{\mathcal{S}}}$ and epistemic uncertainty $\sigma_n : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^{d_{\mathcal{S}}}$ for every $n = 1, \ldots, N$, using transitions from all previously collected trajectories $\mathcal{D}_{1:n}$. This type of statistical models can be learned in practice via various (approximate) Bayesian inference methods [e.g., 57, 58, 59]. Additionally, calibration can be achieved empirically with post-hoc recalibration techniques [60].

Pessimistic policy priors. The set $\mathcal{F}_0$ has a unique interpretation: it formally represents the agent's prior knowledge about the environment. For instance, in fully data-driven settings, given an offline dataset $\mathcal{D}_0$, a model $\mathcal{F}_0$ can be learned from $\mathcal{D}_0$ and then used to construct a conservative MDP that penalizes the agent when it goes outside of the data distribution [61, 62]. Alternatively, in simulator-driven workflows, one has access to a nominal model $\mu_0 : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^{d_{\mathcal{S}}}$, and a bounded 'sim-to-real gap' $u : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^{d_{\mathcal{S}}}$, giving $f \in \mathcal{F}_0 = \{\tilde{f} \mid |\tilde{f} - \mu_0| \leq u\}$. Probabilistic models of this sort formalize the conservatism needed for safe prior policies.

Assumption 4 (Safe policy prior)**.** *We assume access to a pessimistic policy prior $\hat{\pi} : \mathcal{S} \to \mathcal{A}$ that satisfies the constraint in Equation (2) for all dynamics $f \in \mathcal{F}_0$ such that under the true dynamics we have $V_c^{\hat{\pi}}(s) \leq V_c^{\pi_c^*}(s)$ for all states with probability at least $1 - \delta$.*

Learning such "worst-case" policies builds on the principle of pessimism in the face of uncertainty and has been extensively studied in the robust RL literature [63, 64, 65, 66], offline safe RL [67, 68, 69, 70], safe learning from demonstrations [71, 72] and when transferring from simulators [73, 74]. In our experiments in Section 5, we leverage a few of these methods and demonstrate the effectiveness of combining them with safe online learning.

4 SOOPER: Safe Online Optimism for Pessimistic Expansion in RL

In a nutshell, SOOPER can be explained by two modes of operation: **(i)** safe online data collection by pessimistically invoking a safe prior policy early enough so as to ensure constraint satisfaction; **(ii)** optimistic planning in "simulation" via model-based rollouts, driving exploration of the unknown dynamics through intrinsic rewards.

Quantifying pessimism. We define the pessimistic cost value function $\bar{V}_c^{\hat{\pi}}$ for the policy prior $\hat{\pi}$:

$$\bar{V}_c^{\hat{\pi}}(s) := \max_{\tilde{f} \in \mathcal{F}_n} \mathbb{E}_{\hat{\pi}}\left[\sum_{t=0}^{\infty} \gamma^t c(s_t, a_t) \,\middle|\, s_0 = s\right], \quad s_{t+1} = \tilde{f}(s_t, a_t) + \omega_t. \tag{4}$$

By definition, Equation (4) considers the worst-case model $\tilde{f}$ in the well-calibrated set $\mathcal{F}_n$ (per Definition 1), which implies that $\bar{V}_c^{\hat{\pi}}(s)$ upper-bounds $V_c^{\hat{\pi}}(s)$ for all $s \in \mathcal{S}$ with probability $1 - \delta$. We later formally show that, as the number of episodes n increases, the agent refines its model $\mathcal{F}_n$ and tightens the pessimistic upper-confidence bound $\bar{V}_c^{\hat{\pi}}$ on the expected accumulated cost.

A tractable upper bound. Computing the pessimistic cost-value $\bar{V}_c^{\hat{\pi}}$ in Equation (4) is intractable due to the maximization over $\mathcal{F}_n$. Instead, we upper-bound $\bar{V}_c^{\hat{\pi}}$ by augmenting the cost function with an uncertainty penalty such that

$$\begin{aligned} &Q_{c,n}^{\hat{\pi}}(s, a) := \mathbb{E}_{\hat{\pi}}\left[\sum_{t=0}^{\infty} \gamma^t \left(c(s_t, a_t) + \lambda_{\text{pessimism}} \|\sigma_n(s_t, a_t)\|\right) \,\middle|\, s_0 = s, a_0 = a\right], \\ &\bar{V}_c^{\hat{\pi}}(s) \leq V_{c,n}^{\hat{\pi}}(s) = \mathbb{E}_{a \sim \hat{\pi}}\left[Q_{c,n}^{\hat{\pi}}(s, a)\right], \; s_{t+1} = \mu_n(s_t, a_t) + \omega_t. \end{aligned} \tag{5}$$

Crucially, directly penalizing the cost function avoids finding the worst-case model within $\mathcal{F}_n$, and lends itself gracefully to existing temporal difference techniques for value function learning. The derivation of $\lambda_{\text{pessimism}}$ and the proof for this upper bound are deferred to Lemma 2 in Appendix A.

Online cost tracking. SOOPER operates in an online fashion, where for each of the trajectory rollouts, it uses $Q_{c,n}^{\hat{\pi}}$ to predict when $a_t \sim \pi_n(\cdot|s_t)$ is potentially unsafe. Specifically, this may occur if the sum of the accumulated cost until t, i.e., $c_{<t} = \sum_{\tau=0}^{t-1} \gamma^\tau c(s_\tau, a_\tau)$ and the pessimistic cost value $\gamma^t Q_{c,n}^{\hat{\pi}}(s_t, a_t)$ exceeds the safety budget d. We exploit this idea in our design of Algorithm 1. In Theorem 1 we show that this design guarantees safety throughout all episodes $n = 1, \ldots, N$.

Algorithm 1 Safe Rollout via Online Cost Tracking

```
Require: Pessimistic cost value $Q_{c,n}^{\hat{\pi}}$, policy $\pi_n$, safe policy prior $\hat{\pi}$, and safety budget $d$
Initialize accumulated cost $c_{<0} = 0$ and experience buffer $\mathcal{D}_n = \emptyset$
for step $t = 1, \ldots, T_n$ do
    Execute $a_t \sim \bar{\pi}_n(a_t|s_t, c_{<t}, Q_{c,n}^{\hat{\pi}})$, observe $s_t, a_t, s_{t+1}$        ▷ Theorem 1
    Append the tuple $(s_t, a_t, s_{t+1})$ to $\mathcal{D}_n$
    Update accumulated cost: $c_{<t+1} = c_{<t} + \gamma^t c_t$
end for
return $\mathcal{D}_n$
```

Theorem 1 (Safety guarantee)**.** *Suppose Assumptions 1 to 4 hold and $\mathcal{F}_n$ is well-calibrated $\forall n = 1, \ldots, N$ according to Definition 1. If actions are executed for all timesteps t according to*

$$\bar{\pi}_n(a_t|s_t, c_{<t}, Q_{c,n}^{\hat{\pi}}) := \begin{cases} \pi_n(\cdot|s_t) & \textit{if } \Phi(s_t, a_t, c_{<t}, Q_{c,n}^{\hat{\pi}}) < d \\ \hat{\pi}(s_t) & \textit{otherwise,} \end{cases}$$

where Φ is the discounted sum of the realized accumulated cost $c_{<t} := \sum_{\tau=0}^{t-1} \gamma^\tau c(s_\tau, a_\tau)$ until $t-1$ and the pessimistic cost value $Q_{c,n}^{\hat{\pi}}$ such that $\Phi(s_t, a_t, c_{<t}, Q_{c,n}^{\hat{\pi}}) := c_{<t} + \gamma^t Q_{c,n}^{\hat{\pi}}(s_t, a_t)$ with factor $\lambda_{\textit{pessimism}} = \bar{C} \frac{\gamma}{1-\gamma} \frac{(1+\sqrt{d_{\mathcal{S}}})\beta_N(\delta)}{\sigma}$, where $\bar{C} = \max\{C_{\textit{max}}, k_{\textit{max}}\}$. Then, Algorithm 1 satisfies the safety constraint with probability $1-\delta$ on f for every episode $n = 1, \ldots, N$.

We refer to Appendix A for the formal proof. Monitoring costs online allows the agent to execute π_n until $\hat{\pi}$ is required. This enables the agent to visit state-actions that $\hat{\pi}$ would avoid, since those would be deemed unsafe by $\bar{V}_c^{\hat{\pi}}(s_0)$.

Simulated exploration. While Theorem 1 guarantees safety during *online* rollouts, the policy π_n may explore potentially unsafe behaviors by using simulated rollouts on the learned world model. This allows SOOPER to train π_n on an *unconstrained* "planning MDP" $\widetilde{\mathcal{M}}$, that differs from the real CMDP $\mathcal{M}_c$ in that it reaches a terminal state $s_\dagger$ after those state-actions that would trigger $\hat{\pi}$ during online deployment:

$$\tilde{p}(s_{t+1} \mid s_t, a_t) := \begin{cases} \mathbf{1}\{s_{t+1} = s_\dagger\} & \text{if } \Phi(a_t, s_t, c_{<t}, Q_{c,n}^{\hat{\pi}}) \geq d \text{ or } s_t = s_\dagger, \\ p(s_{t+1} \mid s_t, a_t) & \text{otherwise.} \end{cases} \tag{6}$$

The resulting termination signal encourages the agent to find policies that outperform $\hat{\pi}$ if possible. This is done by assigning the terminal reward on $\widetilde{\mathcal{M}}$ to the pessimistic value $\underline{V}_r^{\hat{\pi}}$ of $\hat{\pi}$ upon termination:

$$\tilde{r}(s_t, a_t) := \begin{cases} \underline{V}_r^{\hat{\pi}}(s_t) & \text{if } \Phi(a_t, s_t, c_{<t}, Q_{c,n}^{\hat{\pi}}) \geq d, \\ 0 & \text{if } s_t = s_\dagger, \\ r(s_t, a_t) & \text{otherwise,} \end{cases} \tag{7}$$

whereby,

$$\underline{V}_r^{\hat{\pi}}(s) := \min_{\tilde{f} \in \mathcal{F}_n} \mathbb{E}_{\hat{\pi}} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \,\middle|\, s_0 = s \right], \quad s_{t+1} = \tilde{f}(s_t, a_t) + \omega_t. \tag{8}$$

This choice of the terminal reward keeps the reward structure consistent with the rewards encountered during online rollouts of $\hat{\pi}$ on f, unlike prior works that treat the terminal reward as a hyperparameter [36, 8]. Figure 1 illustrates a concrete example of how this design encourages the agent to learn to outperform the prior policy.

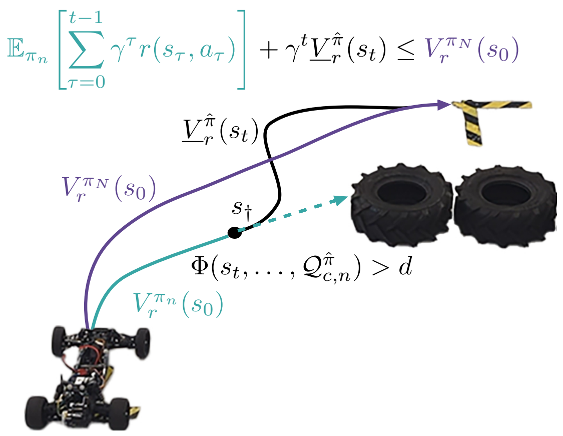

$\mathbb{E}_{\pi_n}\left[\sum_{\tau=0}^{t-1}\gamma^\tau r(s_\tau, a_\tau)\right] + \gamma^t \underline{V}_r^{\hat{\pi}}(s_t) \leq V_r^{\pi_N}(s_0)$
$\underline{V}_r^{\hat{\pi}}(s_t)$
$V_r^{\pi_N}(s_0)$
$s_\dagger$
$\Phi(s_t, \dots, \mathcal{Q}_{c,n}^{\hat{\pi}}) > d$
$V_r^{\pi_n}(s_0)$

Figure 1: The agent's goal is to reach the cross marker while avoiding obstacles (tires). The agent deploys a policy π_n, however at time t, it switches to the prior policy $\hat{\pi}$, to maintain the safety criterion of Algorithm 1. The prior policy $\hat{\pi}$ ensures safety by following a conservative route, but sacrifices performance, resulting in a lower return $\underline{V}_r^{\hat{\pi}}(s_t)$. The trajectory of iteration n is recorded to improve models of subsequent iterations. After N iterations, as more data is gathered, the agent learns a more rewarding trajectory via model-generated rollouts in $\widetilde{\mathcal{M}}$. See provided video.

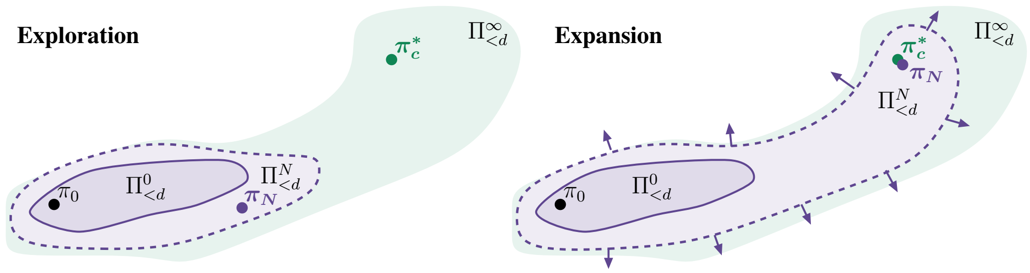


Figure 2: We denote the implicit set of safe policies in iteration n by $\Pi^n_{<d}$, based on the learned model $\mathcal{F}_n$. **Left:** exploration-exploitation in constrained tasks may not find an optimal policy because search is limited to $\Pi^n_{<d}$. **Right:** expansion proactively enlarges the safe set and reaches the optimum.

Exploration-exploitation and expansion. Probabilistic models have proven particularly effective in the unconstrained setting. Optimistic planning with these models enables a principled balance between exploration and exploitation, forming the basis of the celebrated UCRL algorithm [75, 56]. However, when safety must hold at every iteration n, exploration-exploitation is restricted to the set of policies that are safe with high probability. In particular, this set consists only those policies that satisfy the constraint under the *worst-case* model in $\mathcal{F}_n$. Crucially, due to pessimism, an optimal policy π_c^* may not lie in this set initially, as shown in Figure 2. Nonetheless, this set can be *expanded* by updating the model using the collected measurements from the environment, even in those regions where rewards are low. Prior works typically follow an expand-then-explore-exploit strategy [12, 16, 30, 9], which begins by learning the dynamics with reward-free trajectory sampling to expand the safe set, followed by exploration-exploitation to find an optimal and safe policy. This approach has two limitations: **(i)** it requires agents to first learn an auxiliary policy unrelated to the task, leading to wasted compute and exploration; **(ii)** it restricts the theoretical analysis to simple regret, allowing arbitrarily poor performance during exploration as long as the final policy is near-optimal.

A unified objective. In contrast, SOOPER addresses the exploration-exploitation-expansion dilemma through *intrinsic rewards* [76] with a single tractable objective:

$$\begin{aligned}\pi_n \in \arg\max_{\pi} \mathbb{E}_{\tilde{p},\pi,s_0}\left[\sum_{t=0}^{\infty}\gamma^t \tilde{r}(s_t, a_t) + (\gamma^t \lambda_{\text{explore}} + \sqrt{\gamma^t}\lambda_{\text{expand}})\|\sigma_n(s_t, a_t)\|\right], \\ s_{t+1} = \mu_n(s_t, a_t) + \omega_t, \quad a_t \sim \pi(\cdot|s_t),\end{aligned} \tag{9}$$

where λ_{explore} and λ_{expand} are weighting factors for the respective components of the intrinsic reward bonus, encouraging exploration and expansion. In Lemma 12 in Appendix B, we derive these quantities in closed form and prove that Equation (9) indeed enables expansion. Importantly, since we only augment the reward with exploration bonuses, an approximate solution to Equation (9) can be computed efficiently.

Converging to the optimum. By iteratively updating the model $\mathcal{F}_n$ and planning according to Equation (9) in each episode, SOOPER achieves *sublinear cumulative regret* and therefore converges for a sufficient number of episodes N in Algorithm 2 according to Theorem 2.

Algorithm 2 Safe Online Optimism for Pessimistic Expansion in RL (SOOPER)

```
Require: Initial dynamics set $\mathcal{F}_0$, pessimistic cost value $Q^{\hat{\pi}}_{c,0}$, policy prior $\hat{\pi}$, safety budget $d$
1: Initialize experience buffer $\mathcal{D}$
2: for episode $n = 1, \ldots, N$ do
3:     Execute Algorithm 1 to collect $\mathcal{D}_n$                ▷ On the real environment
4:     Update $\mathcal{F}_n \leftarrow \mathcal{D}_{1:n}$
5:     Obtain $\pi_n, Q^{\hat{\pi}}_{c,n}$ via Equations (5) and (9)    ▷ Model-based exploration and planning
6: end for
7: return $\pi_n$
```

Theorem 2 (Sublinear cumulative regret)**.** *Suppose Assumptions 1 to 4 hold and the model $\mathcal{F}_n$ is well calibrated according to Definition 1. Then, Algorithm 2 guarantees with probability $1 - \delta$*

$$R(N) = \sum_{n=1}^{N} \left(J_r(\pi_c^*, f) - J_r(\bar{\pi}_n, f) \right) \leq \mathcal{O}\left(\Gamma^{7/2}_{N \log(N)} \sqrt{N} \right),$$
$$\textit{and } J_c(\bar{\pi}_n, f) \leq d, \ \forall n \in \{1, \ldots, N\},$$

that is, the cumulative regret grows sub-linearly in N, while satisfying the constraint throughout learning by Theorem 1.

Proof sketch. Previous works explicitly expand the safe set of policies via pure exploration. This restricts them to a simple regret analysis, since performance can be evaluated *only after* the reward-free expansion phase. SOOPER optimizes a single objective, thus expansion occurs *implicitly, during learning the task*. This allows us to extend the analysis to cumulative regret, providing guarantees on the agent's performance throughout learning. We bound the cumulative regret in Equation (3) by decomposing the performance gap in each iteration n into two terms (Lemma 4). **(i)** The first term corresponds to the regret on the planning MDP $\widetilde{\mathcal{M}}$ (cf. Equations (6) and (7)). Specifically, we compare the performance difference between an optimal policy of $\widetilde{\mathcal{M}}$ and the learned policy π_n. Since this setting is purely unconstrained, we bound this term following the regret analysis for unconstrained MDPs of Kakade et al. [53]. **(ii)** Next, we analyze the performance gap due to invoking the prior policy $\hat{\pi}$ (Algorithm 1). In particular, we study the performance of executing π_c^*, an optimal policy of the true CMDP $\mathcal{M}_c$, however under the safety criterion of Algorithm 1. Intuitively, due to the pessimism of $Q^{\hat{\pi}}_{c,n}$, instead of performing optimally with π_c^*, Algorithm 1 invokes $\hat{\pi}$, thus leading to suboptimal performance. The gap in performance between executing π_c^* freely and under Algorithm 1 is bounded by the model uncertainty at iteration n (Lemmas 6 and 8). The key intuition is that since π_c^* is in fact safe, as model uncertainty shrinks, Algorithm 1 needs to invoke $\hat{\pi}$ less, hence converging to optimal performance. Finally, we show that the sum of these two terms, is upper-bounded by the objective in Equation (9) in each iteration n. The complete proof is detailed in Appendix B.

Practical implementation. Our practical implementation follows closely Algorithm 2. We instantiate a "deep" version of Algorithm 2 that scales to high-dimensional continuous control tasks, following a model-based actor-critic architecture similar to MBPO [77]. Line 4 is implemented via a probabilistic ensemble of neural-networks [78]. In our practical implementation we additionally learn the reward and cost functions (see Appendix C). The standard deviation across ensemble predictions is used to estimate the epistemic uncertainty σ_n [79]. We found this approach to deliver reliable uncertainty estimates in practice (cf. Section 5 and Appendix D). In Line 5, we solve Equations (5) and (9) using model-generated rollouts and by performing a fixed number of actor-critic updates as described by Janner et al. [77]. Adapting MBPO to our setting is straightforward, as it only requires wrapping the model predictions within the MDP defined by $\widetilde{\mathcal{M}}$. This suggests that SOOPER is readily extensible to future advancements in deep RL.

5 Experiments

We now present our empirical results, demonstrating SOOPER's safety and performance guarantees in practice. In the Appendix D, we provide additional ablations, illustrating how each component of SOOPER plays a key role in attaining good performance while maintaining safety throughout learning.

Setup. In all experiments, we use the model-based architecture described in Section 4. We compare SOOPER with the following baselines: **(i)** SAILR [8], a state-of-the-art RL algorithm for safe

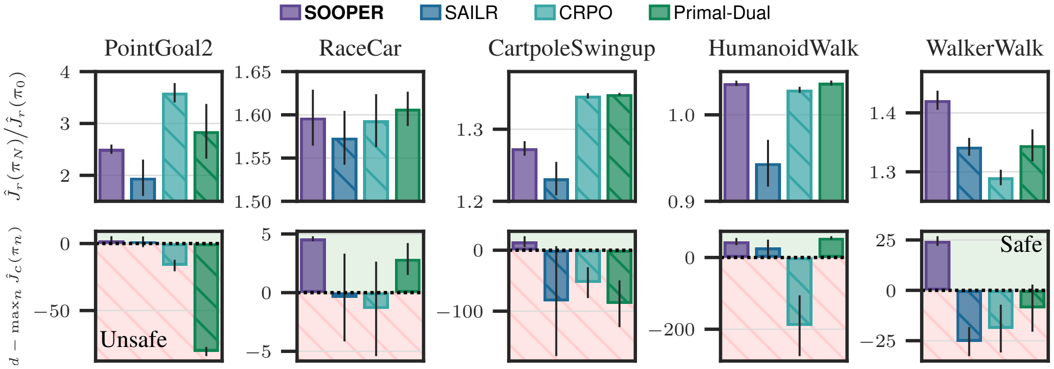


Figure 3: Performance improvement over the baseline policy and largest constraint recorded throughout learning. Among all methods, SOOPER remains safe in all tasks while consistently outperforming or being on par with the other baselines when they satisfy the constraints.

exploration. **(ii)** CRPO [80], a CMDP solver without safety guarantees during learning. **(iii)** Similarly, Primal-Dual [81] is a constrained optimization algorithm, which is commonly used as a CMDP solver. CRPO and Primal-Dual serve as natural baselines, as they represent the standard approaches practitioners would typically adopt when fine-tuning policies initialized from simulation or offline training. Unless specified otherwise, we use a batch of 128 trajectories to obtain empirical estimates of the undiscounted cumulative rewards $\hat{J}_r(\pi)$ and costs $\hat{J}_c(\pi)$ and run each experiment for five random seeds. We slightly abuse notation, so that for SOOPER, π denotes the policy induced by Algorithm 1 and for SAILR the policy induced by their "advantage-based intervention". All baselines are initialized with the same task-specific prior policy.

Safe transfer under dynamics mismatch. We study policies that are trained under a mismatch in the dynamics on five tasks from RWRL [10], SafetyGym [11] and RaceCar [82]. We provide additional details in Appendices E to G about how mismatches in the dynamics are implemented for each task. We refer to Appendix D for additional details on how we train the pessimistic prior policies. We train prior policies under the shifted dynamics (μ_0, cf. Section 3) and continue learning on the true dynamics. Figure 3 presents the largest cumulative cost recorded on the true dynamics throughout training, along with the relative performance improvement at the end of training compared to the policy prior before online learning. As shown, SOOPER satisfies the constraints throughout learning in all of the tasks. For each task, among all algorithms that satisfy the constraints, SOOPER is consistently on par or outperforms the baselines. This demonstrates SOOPER's ability to transfer safely under distribution shifts.

Safe offline-to-online. We evaluate SOOPER when initialized with a conservative policy using only offline data. To this end, we collect a fixed offline data of 2M transitions from the PointGoal1 task of SafetyGym. We train a pessimistic prior policy offline by following MOPO [62], a well-established model-based offline RL algorithm. We use a Primal-Dual optimizer to solve the constraint optimization problem with this offline data. We found this approach to yield policies that satisfy the constraint while still delivering nontrivial performance. We use these policies to initialize all the baselines for our experiments. Figure 5a presents the objective and constraint during online learning. Notably, SOOPER outperforms all the baselines while maintaining safety.

Scaling to vision control. Next, we demonstrate that SOOPER scales to control tasks with image-based observations. Specifically, we use the CartpoleSwingup task from Dulac-Arnold et al. [10] and train policies that operate on three temporally-stacked 64×64 grayscale images as depicted in Figure 4. As in Figure 3, we introduce a distribution shift in the dynamics and pretrain a vision policy under the perturbed dynamics using DrQ [83]. We then train SOOPER by learning the dynamics model directly on the embeddings produced by the pretrained DrQ vision encoder. Figure 5b shows that even when trained on these embeddings, SOOPER satisfies the safety constraints and achieves near-optimal performance.

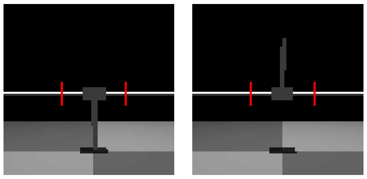
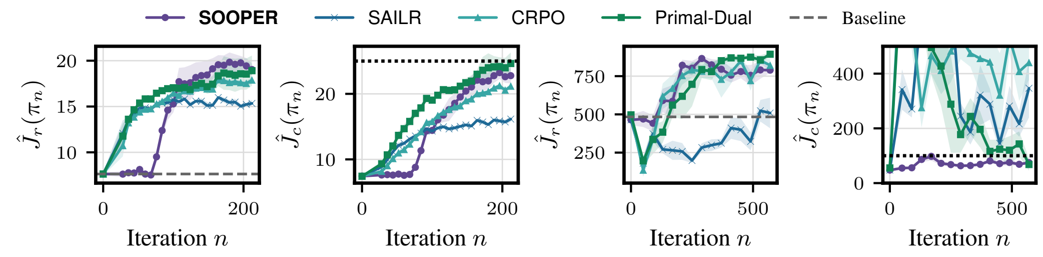


Figure 4: Image observations for the CartpoleSwingup task. The goal is to swing the pendulum to the upright position while avoiding the range outside the vertical red lines.



(a) Safe online learning from an offline-trained policy.

(b) Vision control task.

Figure 5: Learning curves of the objective and constraint when learning from offline and vision policies. SOOPER satisfies the constraints while significantly improving over the initial prior policy.

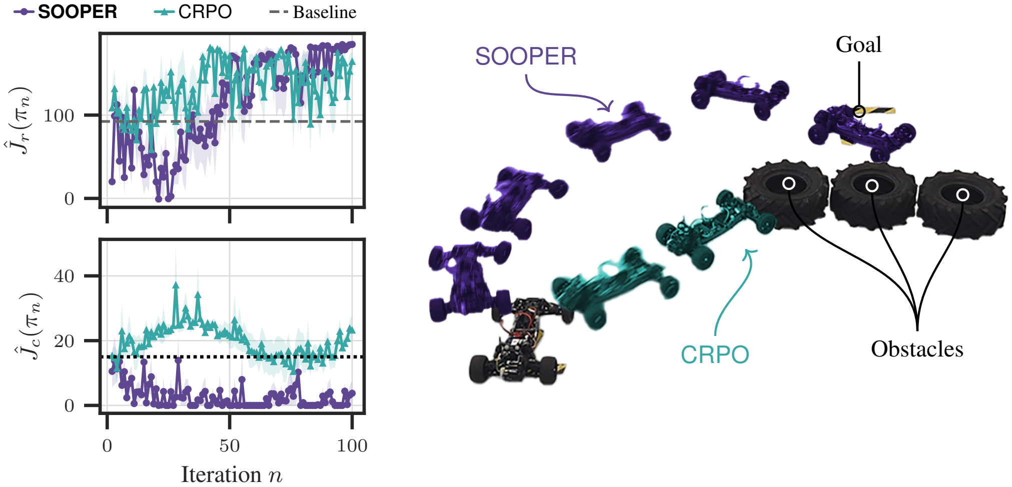


Figure 6: Safe exploration on real hardware with SOOPER. We report the mean and standard error across five seeds of the objective and constraint measured on the real system. SOOPER learns to improve over the prior policy while satisfying the constraints throughout learning. Video of training.

Safe online learning on real hardware. Finally, we evaluate SOOPER when trained online *directly on real-world robotic hardware*: a highly dynamic, remote-controlled race car operating at 60 Hz. The task is to reach a designated goal position located behind three obstacles (tires) that must be avoided (illustrated in Figure 6). This setting is particularly challenging due to stochasticity introduced by high-frequency control and delays in both actuation and motion-capture measurements. We initialize training with a prior policy derived from a first-principles simulator [82], and then fine-tune it on the physical system using trajectories collected in real time. We repeat the experiment with five random seeds and report cumulative rewards and costs after each iteration in Figure 6. As shown, after training, SOOPER is roughly twice as good as the prior policy in terms of rewards while satisfying the safety constraint throughout learning. In Appendix D, we extend this experiment to the offline setting, and report results when using a prior policy trained offline with data collected from the real system.

6 Conclusion

In this work, we address a critical challenge in systems that learn autonomously, namely, maintaining safety during learning. We present SOOPER, a novel algorithm that provably improves performance over policies trained in simulation or offline, while satisfying safety constraints throughout learning. SOOPER is simple to implement, achieves state-of-the-art performance and consistently maintains safety when deployed in practice. In addition, we establish a new theoretical result on the cumulative regret, enabling a guarantee of good performance during learning and not only at the last iteration. These contributions advance our understanding of safe exploration and make an important step towards deployable RL. Nonetheless, many important challenges remain. In terms of theory and methodology, extending our work to high-probability constraints over particular states and developing new methods for the non-episodic setting, where agents learn from a single trajectory without resets, are two challenging yet crucial problems. On the practical side, considering more complex tasks with more powerful models is a promising direction for future work.

References

[1] David Silver and Richard S Sutton. Welcome to the era of experience. *Google AI*, 2025. (Cited on page 1)

[2] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of Go without human knowledge. *Nature*, 2017. (Cited on page 1)

[3] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022. (Cited on page 1)

[4] David Dalrymple, Joar Skalse, Yoshua Bengio, Stuart Russell, Max Tegmark, Sanjit Seshia, Steve Omohundro, Christian Szegedy, Ben Goldhaber, Nora Ammann, et al. Towards guaranteed safe ai: A framework for ensuring robust and reliable ai systems. *arXiv preprint arXiv:2405.06624*, 2024. (Cited on page 1)

[5] Shane Legg. System two safety. The Alignment Workshop, 2023. (Cited on page 1)

[6] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety, 2016. (Cited on pages 1 and 3)

[7] E. Altman. *Constrained Markov Decision Processes*. Chapman and Hall, 1999. (Cited on pages 1 and 3)

[8] Nolan Wagener, Byron Boots, and Ching-An Cheng. Safe reinforcement learning using advantage-based intervention. In *International Conference on Machine Learning*, 2021. (Cited on pages 2, 3, 5, and 7)

[9] Yarden As, Bhavya Sukhija, Lenart Treven, Carmelo Sferrazza, Stelian Coros, and Andreas Krause. Actsafe: Active exploration with safety constraints for reinforcement learning. In *International Conference on Learning Representations*, 2025. (Cited on pages 2, 6, and 20)

[10] Gabriel Dulac-Arnold, Nir Levine, Daniel J. Mankowitz, Jerry Li, Cosmin Paduraru, Sven Gowal, and Todd Hester. An empirical investigation of the challenges of real-world reinforcement learning, 2021. (Cited on pages 2, 8, and 38)

[11] Alex Ray, Joshua Achiam, and Dario Amodei. Benchmarking Safe Exploration in Deep Reinforcement Learning. 2019. (Cited on pages 2, 8, and 38)

[12] Yanan Sui, Alkis Gotovos, Joel Burdick, and Andreas Krause. Safe exploration for optimization with gaussian processes. In *International Conference on Machine Learning*, 2015. (Cited on pages 2 and 6)

[13] Felix Berkenkamp, Andreas Krause, and Angela P. Schoellig. Bayesian optimization with safety constraints: safe and automatic parameter tuning in robotics. *Machine Learning*, 2016. (Cited on page 2)

[14] Johannes Kirschner, Mojmir Mutny, Nicole Hiller, Rasmus Ischebeck, and Andreas Krause. Adaptive and safe Bayesian optimization in high dimensions via one-dimensional subspaces. In *International Conference on Machine Learning*, 2019. (Cited on page 2)

[15] Bhavya Sukhija, Matteo Turchetta, David Lindner, Andreas Krause, Sebastian Trimpe, and Dominik Baumann. Gosafeopt: Scalable safe exploration for global optimization of dynamical systems. *Artificial Intelligence*, 2023. (Cited on page 2)

[16] Felix Berkenkamp, Matteo Turchetta, Angela P. Schoellig, and Andreas Krause. Safe model-based reinforcement learning with stability guarantees. In *International Conference on Neural Information Processing Systems*, 2017. (Cited on pages 2, 3, and 6)

[17] Manish Prajapat, Johannes Köhler, Matteo Turchetta, Andreas Krause, and Melanie N. Zeilinger. Safe guaranteed exploration for non-linear systems. *IEEE Transactions on Automatic Control*, 2025. (Cited on page 2)

[18] Torsten Koller, Felix Berkenkamp, Matteo Turchetta, and Andreas Krause. Learning-based model predictive control for safe exploration. *IEEE Conference on Decision and Control*, 2018. (Cited on page 2)

[19] Manish Prajapat, Amon Lahr, Johannes Köhler, Andreas Krause, and Melanie N Zeilinger. Towards safe and tractable gaussian process-based mpc: Efficient sampling within a sequential quadratic programming framework. In *IEEE Conference on Decision and Control*. IEEE, 2024. (Cited on page 2)

[20] Manish Prajapat, Johannes Köhler, Amon Lahr, Andreas Krause, and Melanie N Zeilinger. Finite-sample-based reachability for safe control with gaussian process dynamics. *arXiv preprint arXiv:2505.07594*, 2025. (Cited on page 2)

[21] Lukas Hewing, Juraj Kabzan, and Melanie N Zeilinger. Cautious model predictive control using gaussian process regression. *IEEE Transactions on Control Systems Technology*, 2019. (Cited on page 2)

[22] Kim Peter Wabersich and Melanie N Zeilinger. A predictive safety filter for learning-based control of constrained nonlinear dynamical systems. *Automatica*, 2021. (Cited on page 2)

[23] Sebastian Curi, Armin Lederer, Sandra Hirche, and Andreas Krause. Safe reinforcement learning via confidence-based filters. In *Conference on Decision and Control*. IEEE, 2022. (Cited on pages 2 and 3)

[24] Teodor Mihai Moldovan and Pieter Abbeel. Safe exploration in markov decision processes. *arXiv preprint arXiv:1205.4810*, 2012. (Cited on page 2)

[25] Matteo Turchetta, Felix Berkenkamp, and Andreas Krause. Safe exploration in finite markov decision processes with gaussian processes. *International Conference on Neural Information Processing Systems*, 2016. (Cited on page 2)

[26] Akifumi Wachi, Yanan Sui, Yisong Yue, and Masahiro Ono. Safe exploration and optimization of constrained mdps using gaussian processes. *AAAI*, 2018. (Cited on page 2)

[27] Akifumi Wachi and Yanan Sui. Safe reinforcement learning in constrained markov decision processes. In *International Conference on Machine Learning*, 2020. (Cited on page 2)

[28] Thiago D Simão, Nils Jansen, and Matthijs TJ Spaan. Alwayssafe: Reinforcement learning without safety constraint violations during training. 2021. (Cited on page 2)

[29] Manish Prajapat, Matteo Turchetta, Melanie Zeilinger, and Andreas Krause. Near-optimal multi-agent learning for safe coverage control. *International Conference on Neural Information Processing Systems*, 2022. (Cited on page 2)

[30] Archana Bura, Aria HasanzadeZonuzy, Dileep Kalathil, Srinivas Shakkottai, and Jean-Francois Chamberland. Dope: Doubly optimistic and pessimistic exploration for safe reinforcement learning. *International Conference on Neural Information Processing Systems*, 2022. (Cited on pages 2 and 6)

[31] Gal Dalal, Krishnamurthy Dvijotham, Matej Vecerik, Todd Hester, Cosmin Paduraru, and Yuval Tassa. Safe exploration in continuous action spaces. *arXiv preprint arXiv:1801.08757*, 2018. (Cited on page 2)

[32] Benjamin Eysenbach, Shixiang Gu, Julian Ibarz, and Sergey Levine. Leave no trace: Learning to reset for safe and autonomous reinforcement learning. In *International Conference on Learning Representations*, 2018. (Cited on page 2)

[33] Krishnan Srinivasan, Benjamin Eysenbach, Sehoon Ha, Jie Tan, and Chelsea Finn. Learning to be safe: Deep rl with a safety critic. *arXiv preprint arXiv:2010.14603*, 2020. (Cited on page 2)

[34] Brijen Thananjeyan, Ashwin Balakrishna, Suraj Nair, Michael Luo, Krishnan Srinivasan, Minho Hwang, Joseph E. Gonzalez, Julian Ibarz, Chelsea Finn, and Ken Goldberg. Recovery rl: Safe reinforcement learning with learned recovery zones. *IEEE Robotics and Automation Letters*, 2021. (Cited on page 2)

[35] Homanga Bharadhwaj, Aviral Kumar, Nicholas Rhinehart, Sergey Levine, Florian Shkurti, and Animesh Garg. Conservative safety critics for exploration. In *International Conference on Learning Representations*, 2021. (Cited on page 2)

[36] Garrett Thomas, Yuping Luo, and Tengyu Ma. Safe reinforcement learning by imagining the near future. In *International Conference on Neural Information Processing Systems*, 2021. (Cited on pages 2 and 5)

[37] Hao Sun, Ziping Xu, Meng Fang, Zhenghao Peng, Jiadong Guo, Bo Dai, and Bolei Zhou. Safe exploration by solving early terminated mdp. *arXiv preprint arXiv:2107.04200*, 2021. (Cited on page 2)

[38] Zuxin Liu, Zhepeng Cen, Vladislav Isenbaev, Wei Liu, Steven Wu, Bo Li, and Ding Zhao. Constrained variational policy optimization for safe reinforcement learning. In *International Conference on Machine Learning*, 2022. (Cited on page 2)

[39] Yarden As, Ilnura Usmanova, Sebastian Curi, and Andreas Krause. Constrained policy optimization via bayesian world models. In *International Conference on Learning Representations*, 2022. (Cited on page 2)

[40] Aivar Sootla, Alexander I Cowen-Rivers, Taher Jafferjee, Ziyan Wang, David H Mguni, Jun Wang, and Haitham Ammar. Saute rl: Almost surely safe reinforcement learning using state augmentation. In *International Conference on Machine Learning*, 2022. (Cited on page 2)

[41] Weidong Huang, Jiaming Ji, Chunhe Xia, Borong Zhang, and Yaodong Yang. Safedreamer: Safe reinforcement learning with world models. In *International Conference on Learning Representations*, 2024. (Cited on page 2)

[42] Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *International Conference on Machine Learning*, 2017. (Cited on page 2)

[43] Nikola Milosevic, Johannes Müller, and Nico Scherf. Embedding safety into rl: A new take on trust region methods. *arXiv preprint arXiv:2411.02957*, 2024. (Cited on page 2)

[44] Yinlam Chow, Ofir Nachum, Edgar Duenez-Guzman, and Mohammad Ghavamzadeh. A lyapunov-based approach to safe reinforcement learning. *International Conference on Neural Information Processing Systems*, 2018. (Cited on page 2)

[45] Yongshuai Liu, Jiaxin Ding, and Xin Liu. Ipo: Interior-point policy optimization under constraints. In *AAAI*, 2020. (Cited on page 2)

[46] Ilnura Usmanova, Yarden As, Maryam Kamgarpour, and Andreas Krause. Log barriers for safe black-box optimization with application to safe reinforcement learning. *Journal of Machine Learning Research*, 2024. (Cited on pages 2 and 35)

[47] Tingting Ni and Maryam Kamgarpour. A safe exploration approach to constrained markov decision processes. In *International Conference on Artificial Intelligence and Statistics*, 2025. (Cited on page 2)

[48] Ilnura Usmanova and Kfir Yehuda Levy. Safe primal-dual optimization with a single smooth constraint. *arXiv preprint arXiv:2505.09349*, 2025. (Cited on page 2)

[49] Akifumi Wachi, Wataru Hashimoto, Xun Shen, and Kazumune Hashimoto. Safe exploration in reinforcement learning: A generalized formulation and algorithms. In *International Conference on Neural Information Processing Systems*, 2023. (Cited on pages 2 and 3)

[50] Javier García, Fern, and o Fernández. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 2015. (Cited on page 3)

[51] Lukas Brunke, Melissa Greeff, Adam W Hall, Zhaocong Yuan, Siqi Zhou, Jacopo Panerati, and Angela P Schoellig. Safe learning in robotics: From learning-based control to safe reinforcement learning. *Annual Review of Control, Robotics, and Autonomous Systems*, 2022. (Cited on page 3)

[52] Shangding Gu, Long Yang, Yali Du, Guang Chen, Florian Walter, Jun Wang, and Alois Knoll. A review of safe reinforcement learning: Methods, theories and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. (Cited on page 3)

[53] Sham Kakade, Akshay Krishnamurthy, Kendall Lowrey, Motoya Ohnishi, and Wen Sun. Information theoretic regret bounds for online nonlinear control. In *International Conference on Neural Information Processing Systems*, 2020. (Cited on pages 3, 7, 18, 20, 23, and 25)

[54] Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014. (Cited on page 3)

[55] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. 2015. (Cited on page 3)

[56] Sebastian Curi, Felix Berkenkamp, and Andreas Krause. Efficient model-based reinforcement learning through optimistic policy search and planning. In *International Conference on Neural Information Processing Systems*, 2020. (Cited on pages 4 and 6)

[57] Carl Edward Rasmussen and Christopher K. I. Wiliams. *Gaussian Processes for Machine Learning*. 2006. (Cited on page 4)

[58] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *International Conference on Neural Information Processing Systems*, 2017. (Cited on page 4)

[59] Qiang Liu and Dilin Wang. Stein variational gradient descent: A general purpose bayesian inference algorithm. *International Conference on Neural Information Processing Systems*, 2016. (Cited on page 4)

[60] Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. Accurate uncertainties for deep learning using calibrated regression. In *International Conference on Machine Learning*, 2018. (Cited on page 4)

[61] Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020. (Cited on page 4)

[62] Tianhe Yu, Garrett Thomas, Lantao Yu, Stefano Ermon, James Y Zou, Sergey Levine, Chelsea Finn, and Tengyu Ma. Mopo: Model-based offline policy optimization. *International Conference on Neural Information Processing Systems*, 2020. (Cited on pages 4, 8, and 34)

[63] Toshinori Kitamura, Tadashi Kozuno, Wataru Kumagai, Kenta Hoshino, Yohei Hosoe, Kazumi Kasaura, Masashi Hamaya, Paavo Parmas, and Yutaka Matsuo. Near-optimal policy identification in robust constrained markov decision processes via epigraph form. *arXiv preprint arXiv:2408.16286*, 2024. (Cited on page 4)

[64] Zhengfei Zhang, Kishan Panaganti, Laixi Shi, Yanan Sui, Adam Wierman, and Yisong Yue. Distributionally robust constrained reinforcement learning under strong duality. In *Reinforcement Learning Conference*. (Cited on page 4)

[65] Laixi Shi, Gen Li, Yuting Wei, Yuxin Chen, and Yuejie Chi. Pessimistic q-learning for offline reinforcement learning: Towards optimal sample complexity. In *International conference on machine learning*. (Cited on page 4)

[66] Laixi Shi and Yuejie Chi. Distributionally robust model-based offline reinforcement learning with near-optimal sample complexity. *Journal of Machine Learning Research*, 2024. (Cited on page 4)

[67] Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline rl? In *International conference on machine learning*, 2021. (Cited on page 4)

[68] Zuxin Liu, Zijian Guo, Haohong Lin, Yihang Yao, Jiacheng Zhu, Zhepeng Cen, Hanjiang Hu, Wenhao Yu, Tingnan Zhang, Jie Tan, et al. Datasets and benchmarks for offline safe reinforcement learning. *arXiv preprint arXiv:2306.09303*, 2023. (Cited on page 4)

[69] Yinan Zheng, Jianxiong Li, Dongjie Yu, Yujie Yang, Shengbo Eben Li, Xianyuan Zhan, and Jingjing Liu. Safe offline reinforcement learning with feasibility-guided diffusion model. In *International Conference on Learning Representations*, 2024. (Cited on page 4)

[70] Akifumi Wachi, Kohei Miyaguchi, Takumi Tanabe, Rei Sato, and Youhei Akimoto. A provable approach for end-to-end safe reinforcement learning. *arXiv preprint arXiv:2505.21852*, 2025. (Cited on page 4)

[71] Andreas Schlaginhaufen and Maryam Kamgarpour. Identifiability and generalizability in constrained inverse reinforcement learning. In *International Conference on Machine Learning*, 2023. (Cited on page 4)

[72] David Lindner, Xin Chen, Sebastian Tschiatschek, Katja Hofmann, and Andreas Krause. Learning safety constraints from demonstrations with unknown rewards. In *International Conference on Artificial Intelligence and Statistics*, 2024. (Cited on page 4)

[73] James Queeney and Mouhacine Benosman. Risk-averse model uncertainty for distributionally robust safe reinforcement learning. In *International Conference on Neural Information Processing Systems*, 2023. (Cited on pages 4 and 38)

[74] Yarden As, Chengrui Qu, Benjamin Unger, Dongho Kang, Max van der Hart, Laixi Shi, Stelian Coros, Adam Wierman, and Andreas Krause. SPiDR: A simple approach for zero-shot safety in sim-to-real transfer. In *International Conference on Neural Information Processing Systems*, 2025. (Cited on pages 4 and 35)

[75] Peter Auer and Ronald Ortner. Logarithmic online regret bounds for undiscounted reinforcement learning. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *International Conference on Neural Information Processing Systems*, 2006. (Cited on page 6)

[76] Bhavya Sukhija, Lenart Treven, Carmelo Sferrazza, Florian Dorfler, Pieter Abbeel, and Andreas Krause. Optimism via intrinsic rewards: Scalable and principled exploration for model-based reinforcement learning. In *7th Robot Learning Workshop: Towards Robots with Human-Level Abilities*, 2025. (Cited on pages 6, 16, 17, 22, 24, 26, and 30)

[77] Michael Janner, Justin Fu, Marvin Zhang, and Sergey Levine. When to trust your model: Model-based policy optimization. *International Conference on Neural Information Processing Systems*, 2019. (Cited on page 7)

[78] Kurtland Chua, Roberto Calandra, Rowan McAllister, and Sergey Levine. Deep reinforcement learning in a handful of trials using probabilistic dynamics models. *International Conference on Neural Information Processing Systems*, 2018. (Cited on page 7)

[79] Stefan Depeweg, Jose-Miguel Hernandez-Lobato, Finale Doshi-Velez, and Steffen Udluft. Decomposition of uncertainty in bayesian deep learning for efficient and risk-sensitive learning. In *International conference on machine learning*, 2018. (Cited on page 7)

[80] Tengyu Xu, Yingbin Liang, and Guanghui Lan. Crpo: A new approach for safe reinforcement learning with convergence guarantee. In *International Conference on Machine Learning*, 2020. (Cited on page 8)

[81] D. Bertsekas. *Nonlinear Programming*. Athena scientific optimization and computation series. Athena Scientific, 2016. (Cited on page 8)

[82] Juraj Kabzan, Miguel I Valls, Victor JF Reijgwart, Hubertus FC Hendrikx, Claas Ehmke, Manish Prajapat, Andreas Bühler, Nikhil Gosala, Mehak Gupta, Ramya Sivanesan, et al. Amz driverless: The full autonomous racing system. *Journal of Field Robotics*, 2020. (Cited on pages 8, 9, and 39)

[83] Denis Yarats, Ilya Kostrikov, and Rob Fergus. Image augmentation is all you need: Regularizing deep reinforcement learning from pixels. In *International Conference on Learning Representations*, 2021. (Cited on page 8)

[84] Bhavya Sukhija, Lenart Treven, Cansu Sancaktar, Sebastian Blaes, Stelian Coros, and Andreas Krause. Optimistic active exploration of dynamical systems. In *International Conference on Neural Information Processing Systems*, 2023. (Cited on pages 25 and 27)

[85] Sayak Ray Chowdhury and Aditya Gopalan. On kernelized multi-armed bandits. In *International Conference on Machine Learning*, 2017. (Cited on pages 29 and 30)

[86] Eric Lehman, Leighton Thomas, and Meyer Albert. *Mathematics for Computer Science*. 2018. (Cited on page 30)

[87] Yuval Tassa, Yotam Doron, Alistair Muldal, Tom Erez, Yazhe Li, Diego de Las Casas, David Budden, Abbas Abdolmaleki, Josh Merel, Andrew Lefrancq, et al. Deepmind control suite. *arXiv preprint arXiv:1801.00690*, 2018. (Cited on page 38)

A Safety Guarantee

First, we derive in Lemma 1 that the pessimistic safety-condition with online cost tracking (Theorem 1) guarantees safety given the true pessimistic cost value function $\bar{V}_c^{\hat{\pi}}$ of Equation (4).

Lemma 1. *Suppose Assumptions 1 to 4 hold and $\mathcal{F}_n$ is well-calibrated $\forall n = 1, \ldots, N$ according to Definition 1. Given the safe prior $\hat{\pi}$, the rollout of the policy π_n satisfies the safety constraint with probability $1 - \delta$ on f for every episode $n = 1, \ldots, N$, if the actions are executed for all timesteps t according to the pessimistic safety-condition:*

$$\tilde{\pi}_n(a_t | s_t, c_{<t}, \mathcal{F}_n) = \begin{cases} \pi_n(\cdot | s_t) & \textit{if } \bar{\Phi}(s_t, a_t, c_{<t}, \mathcal{F}_n) < d \\ \hat{\pi}(s_t) & \textit{otherwise,} \end{cases}$$

where

$$\bar{\Phi}(s_t, a_t, c_{<t}, \mathcal{F}_n) = c_{<t} + \gamma^t c(s_t, a_t) + \gamma^{t+1} \max_{\tilde{f} \in \mathcal{F}_n} \mathbb{E}_\omega \left[\bar{V}_c^{\hat{\pi}}(\tilde{f}(s_t, a_t) + \omega_t) \right], \tag{10}$$

indicates the sum of the accumulated costs $c_{<t} = \sum_{\tau=0}^{t-1} \gamma^t c(s_t, a_t)$ until $t - 1$, the immediate cost at t, and the worst-case pessimistic cost value $\bar{V}_c^{\hat{\pi}}$ defined in Equation (5).

Proof. We prove the safe rollout of π_n for the unknown f by induction.

1. Base Case ($t = 0$): By definition of $\tilde{\pi}_n$, we only execute $a_0 \sim \pi_n(\cdot | s_0)$ if $\bar{\Phi}(s_0, a_0, c_{<0}, \mathcal{F}_n) < d$ with $c_{<0} = 0$ and otherwise the safe prior $\hat{\pi}$. By definition of $\bar{\Phi}$, either a_0 is safe to execute with

$$d > c(s_0, a_0) + \gamma \max_{\tilde{f} \in \mathcal{F}_n} \mathbb{E}_\omega \left[V_c^{\hat{\pi}}(\tilde{f}(s_0, a_0) + \omega_0) \right], \tag{11}$$

   and otherwise executing $a_0 = \hat{\pi}(s_0)$ is safe since $\bar{V}_c^{\hat{\pi}}(s_0) < d$ by Assumption 4.

2. Inductive Step: Given the agent is safe up to timestep $t - 1$, i.e.,

$$c_{<t-1} + \gamma^{t-1} c(s_{t-1}, a_{t-1}) + \gamma^t \max_{\tilde{f} \in \mathcal{F}_n} \mathbb{E}_\omega \left[V_c^{\hat{\pi}}(\tilde{f}(s_{t-1}, a_{t-1}) + \omega_{t-1}) \right] < d, \tag{12}$$

   we show constraint satisfaction for time t, satisfying

$$d > c_{<t} + \gamma^t c(s_t, a_t) + \gamma^{t+1} \mathbb{E}_\omega \left[V_c^{\hat{\pi}}(f(s_t, a_t) + \omega_t) \right]. \tag{13}$$

   Using $\tilde{\pi}_n$, we either execute a_t if Equation (13) is satisfied and therefore is safe by definition. Otherwise, we invoke the prior $\hat{a}_t = \hat{\pi}(s_t)$. Using $c_{<t} = c_{<t-1} + \gamma^{t-1} c(s_{t-1}, a_{t-1})$, we obtain

$$c_{<t} + \gamma^t c(s_t, \hat{a}_t) + \gamma^{t+1} \max_{\tilde{f} \in \mathcal{F}_n} \mathbb{E}_\omega \left[\bar{V}_c^{\hat{\pi}}(\tilde{f}(s_t, \hat{a}_t) + \omega_t) \right] \tag{14}$$

$$= c_{<t-1} + \gamma^{t-1} c(s_{t-1}, a_{t-1}) + \gamma^t c(s_t, \hat{a}_t) + \gamma^{t+1} \max_{\tilde{f} \in \mathcal{F}_n} \mathbb{E}_\omega \left[\bar{V}_c^{\hat{\pi}}(\tilde{f}(s_t, \hat{a}_t) + \omega_t) \right] \tag{15}$$

$$\leq c_{<t-1} + \gamma^{t-1} c(s_{t-1}, a_{t-1}) + \gamma^t \max_{\tilde{f} \in \mathcal{F}_n} \mathbb{E}_\omega \left[V_c^{\hat{\pi}}(\tilde{f}(s_{t-1}, a_{t-1}) + \omega_{t-1}) \right] < d. \tag{16}$$

□

Next, we show that adding the cost penalties in $Q_{c,n}^{\hat{\pi}}$ (Equation (5)) upper-bounds the true cost value, using Sukhija et al. [76, Lemma A.5]:

Lemma 2. *Suppose Assumptions 1 to 3 hold and $\mathcal{F}_n$ is well-calibrated by Definition 1. Given*

$$\begin{aligned} \gamma^{t+1} \max_{\tilde{f} \in \mathcal{F}_n} \mathbb{E}_\omega \left[\bar{V}_c^{\hat{\pi}}(\tilde{f}(s_t, a_t) + \omega_t) \right] &= \mathbb{E}_{\hat{\pi}} \left[\sum_{\tau=t+1}^{\infty} \gamma^\tau c(s_\tau, a_\tau) \right] \\ \text{s.t.} \;\; s_{\tau+1} = \tilde{f}(s_\tau, a_\tau) + \omega_\tau, \quad a_\tau &= \hat{\pi}(s_\tau), \quad s_{t+1} = \tilde{f}(s_t, a_t) + \omega_t, \\ \lambda_{\textit{pessimism}} = \bar{C} \frac{\gamma}{1 - \gamma} &\frac{\left(1 + \sqrt{d_{\mathcal{S}}}\right) \beta_N(\delta)}{\sigma}, \end{aligned}$$

then for all episodes $n = 1, \dots, N$ with probability of $1 - \delta$:

$$\gamma^{t+1} \max_{\tilde{f} \in \mathcal{F}_n} \mathbb{E}_\omega \left[\bar{V}_c^{\hat{\pi}}(\tilde{f}(s_t, a_t) + \omega_t) \right] \leq \gamma^t \lambda_{\mathit{pessimism}} \|\sigma_n(s_t, a_t)\|$$
$$+ \mathbb{E}_{\hat{\pi}} \left[\sum_{\tau=t+1}^{\infty} \gamma^\tau \left(c(s_\tau, a_\tau) + \lambda_{\mathit{pessimism}} \|\sigma_n(s_\tau, a_\tau)\| \right) \right],$$
$$\text{s.t.} \;\; s_{\tau+1} = \mu_n(s_\tau, a_\tau) + \omega_\tau, \quad a_\tau = \hat{\pi}(s_\tau), \quad s_{t+1} = \mu_n(s_t, a_t) + \omega_t.$$

Proof. Let us denote the value following the nominal dynamics μ_n of $\mathcal{F}_n$ as

$$\tilde{V}_c^{\hat{\pi}}(s) = \mathbb{E}_{\hat{\pi}} \left[\sum_{t=0}^{\infty} \gamma^t c(s_t, a_t) \mid s_0 = s \right], \qquad (17)$$
$$\text{s.t.} \;\; s_{t+1} = \mu_n(s_t, a_t) + \omega_t, \quad a_t = \hat{\pi}(s_t),$$

and the upper bound on the cost value following the worst dynamics $\tilde{f} \in \mathcal{F}_n$ by

$$\bar{V}_c^{\hat{\pi}}(s) = \max_{\tilde{f} \in \mathcal{F}_n} V_c^{\hat{\pi}}(s) = \mathbb{E}_{\hat{\pi}} \left[\sum_{t=0}^{\infty} \gamma^t c(s_t, a_t) \mid s_0 = s \right], \qquad (18)$$
$$\text{s.t.} \;\; s_{t+1} = \tilde{f}(s_t, a_t) + \omega_t, \quad a_t = \hat{\pi}(s_t).$$

Following Sukhija et al. [76, Lemma A.5], the difference between these values is bounded by

$$\left| \bar{V}_c^{\hat{\pi}}(s) - \tilde{V}_c^{\hat{\pi}}(s) \right| \leq \lambda_{\text{pessimism}} \mathbb{E}_{\hat{\pi}} \left[\sum_{t=0}^{\infty} \gamma^t \|\sigma_n(s_t, a_t)\| \mid s_0 = s \right], \qquad (19)$$
$$\text{s.t.} \;\; s_{t+1} = \mu_n(s_t, a_t) + \omega_t, \quad a_t = \hat{\pi}(s_t).$$

Bringing $\tilde{V}_c^{\hat{\pi}}(s)$ to the other side and using the definition in Equation (17) results in:

$$\bar{V}_c^{\hat{\pi}}(s) \leq \tilde{V}_c^{\hat{\pi}}(s) + \lambda_{\text{pessimism}} \mathbb{E}_{\hat{\pi}} \left[\sum_{t=0}^{\infty} \gamma^t \|\sigma_n(s_t, a_t)\| \mid s_0 = s \right], \qquad (20)$$
$$= \mathbb{E}_{\hat{\pi}} \left[\sum_{t=0}^{\infty} \gamma^t \left(c(s_t, a_t) + \lambda_{\text{pessimism}} \|\sigma_n(s_t, a_t)\| \right) \mid s_0 = s \right], \qquad (21)$$
$$\text{s.t.} \;\; s_{t+1} = \mu_n(s_t, a_t) + \omega_t, \quad a_t = \hat{\pi}(s_t). \qquad (22)$$

Further, we use this upper bound having $c(s_t, a_t)$ already observed and rewriting

$$\gamma^{t+1} \max_{\tilde{f} \in \mathcal{F}_n} \mathbb{E}_\omega \left[\bar{V}_c^{\hat{\pi}}(\tilde{f}(s_t, a_t) + \omega_t) \right] = \gamma^t \bar{V}_c^{\hat{\pi}}(s_t) - \gamma^t c(s_t, a_t) \qquad (23)$$
$$\overset{\text{(i)}}{\leq} \mathbb{E}_{\hat{\pi}} \left[\sum_{\tau=t}^{\infty} \gamma^\tau \left(c(s_\tau, a_\tau) + \lambda_{\text{pessimism}} \|\sigma_n(s_\tau, a_\tau)\| \right) \right] - \gamma^t c(s_t, a_t) \qquad (24)$$
$$\leq \mathbb{E}_{\hat{\pi}} \left[\sum_{\tau=t+1}^{\infty} \gamma^\tau \left(c(s_\tau, a_\tau) + \lambda_{\text{pessimism}} \|\sigma_n(s_\tau, a_\tau)\| \right) \right] + \gamma^t \lambda_{\text{pessimism}} \|\sigma_n(s_t, a_t)\|, \qquad (25)$$
$$\text{s.t.} \;\; s_{\tau+1} = \mu_n(s_\tau, a_\tau) + \omega_\tau, \quad a_\tau = \hat{\pi}(s_\tau), \quad s_{t+1} = \mu_n(s_t, a_t) + \omega_t, \qquad (26)$$

where step (i) follows from using the upper bound in Equation (21). □

Theorem 1 (Safety guarantee)**.** *Suppose Assumptions 1 to 4 hold and $\mathcal{F}_n$ is well-calibrated $\forall n = 1, \dots, N$ according to Definition 1. If actions are executed for all timesteps t according to*

$$\bar{\pi}_n(a_t | s_t, c_{<t}, Q_{c,n}^{\hat{\pi}}) \coloneqq \begin{cases} \pi_n(\cdot | s_t) & \textit{if } \Phi(s_t, a_t, c_{<t}, Q_{c,n}^{\hat{\pi}}) < d \\ \hat{\pi}(s_t) & \textit{otherwise,} \end{cases}$$

where Φ is the discounted sum of the realized accumulated cost $c_{<t} \coloneqq \sum_{\tau=0}^{t-1} \gamma^\tau c(s_\tau, a_\tau)$ until $t-1$ and the pessimistic cost value $Q_{c,n}^{\hat{\pi}}$ such that $\Phi(s_t, a_t, c_{<t}, Q_{c,n}^{\hat{\pi}}) \coloneqq c_{<t} + \gamma^t Q_{c,n}^{\hat{\pi}}(s_t, a_t)$ with factor $\lambda_{\mathit{pessimism}} = \bar{C} \frac{\gamma}{1-\gamma} \frac{(1+\sqrt{d_{\mathcal{S}}})\beta_N(\delta)}{\sigma}$, where $\bar{C} = \max\{C_{\mathit{max}}, k_{\mathit{max}}\}$. Then, Algorithm 1 satisfies the safety constraint with probability $1 - \delta$ on f for every episode $n = 1, \dots, N$.

Proof. Theorem 1 is proven using Lemmas 1 and 2. Consider the accumulate cost $\bar{\Phi}(s_t, a_t, c_{<t}, \mathcal{F}_n)$ from Lemma 1

$$
\begin{aligned}
\bar{\Phi}(s_t, a_t, c_{<t}, \mathcal{F}_n) =& c_{<t} + \gamma^t c(s_t, a_t) + \gamma^{t+1} \max_{\tilde{f} \in \mathcal{F}_n} \mathbb{E}_\omega \left[\bar{V}_c^{\hat{\pi}}(\tilde{f}(s_t, a_t) + \omega_t) \right] \qquad (27) \\
\overset{\text{(i)}}{\leq}& c_{<t} + \gamma^t (c(s_t, a_t) + \lambda_{\text{pessimism}} \|\sigma_n(s_t, a_t)\|) \\
&+ \mathbb{E}_{\hat{\pi}} \left[\sum_{\tau=t+1}^{\infty} \gamma^\tau \left(c(s_\tau, a_\tau) + \lambda_{\text{pessimism}} \|\sigma_n(s_\tau, a_\tau)\| \right) \right] \qquad (28) \\
\overset{\text{(ii)}}{=}& c_{<t} + \gamma^t Q_{c,n}^{\hat{\pi}}(s_t, a_t) \overset{\text{(iii)}}{=:} \Phi(s_t, a_t, c_{<t}, Q_{c,n}^{\hat{\pi}}) < d. \qquad (29)
\end{aligned}
$$

Step (i) follows from the upper bound on the cost penalties in Lemma 2, and Step (ii) and (iii) follow from the definitions of $Q_{c,n}^{\hat{\pi}}$ and $\Phi(s_t, a_t, c_{<t}, Q_{c,n}^{\hat{\pi}})$ from Equation (5) and lemma 1. □

B Regret Analysis

Overview of proof. Theorem 2 states that the tractable objective in Equation (9) achieves sublinear cumulative regret while satisfying the safety constraint. We start our proof by decomposing the cumulative regret into per-episode regret terms. Using Lemma 3, we lower-bound the performance of the true system $\mathcal{M}_c$ by the performance on the planning MDP $\widetilde{\mathcal{M}}$. This bound allows us to decompose the per-episode regret into two components in Lemma 4 that can be bounded separately: **(i) Safe rollout vs. optimal rollout.** The first regret term captures the performance gap between the optimal policy π_c^* and the same policy executed via Algorithm 1 on $\mathcal{M}_c$. The difference arises from invoking the prior policy $\hat{\pi}$ due to uncertainty in the learned dynamics. We express this gap in Lemma 5 in terms of action differences between π_c^* and the potentially used safe prior $\hat{\pi}$. By Theorem 1, these action differences can be related in Lemma 6 via the safety criterion Φ to the accumulated model uncertainty along potentially unsafe trajectories starting with $\pi_c^*(a_t|s_t)$ at step t, followed by $\hat{\pi}$. Since it may be unsafe to execute them, we instead express their cumulative uncertainty in terms of the corresponding safe trajectories starting with $\hat{\pi}(s_t)$ at step t (Lemma 7). This is achieved using the difference of Gaussians [53, Lemma C.2]. Consequently, we establish in Lemma 8 an upper bound on the first regret term, which depends on the model uncertainty along the safe rollout of π_c^* under Algorithm 1. **(ii) Planning model vs. true dynamics.** The second regret term quantifies the value gap between an optimal policy for the planning MDP $\widetilde{\mathcal{M}}$ (with true dynamics f) and the policy π_n obtained by optimizing the nominal model μ_n in episode n. This gap can be bounded in terms of the model uncertainty evaluated along the optimal policy π^* when it is safely executed using Algorithm 1. By Lemma 12, the tractable objective in Equation (9) is an optimistic upper bound on the best achievable performance across all models $\mathcal{F}_n$. Consequently, Lemma 13 provides a bound on the sum of the two per-episode regret components along the safely rolled-out behavioral policy π_n using Algorithm 1. Summing these bounds over all episodes yields Theorem 2, which states that the objective in Equation (9) attains sublinear cumulative regret.

To decompose $R(N)$, we first establish a lower bound on the performance of $\bar{\pi}_n$, obtained by safely rolling out π_n on $\mathcal{M}_c$ with Algorithm 1 using Φ.

Planning at the lower performance bound. Given the policy π_n on $\widetilde{\mathcal{M}}$, we derive in Lemma 3 the performance bound of the safe policy $\bar{\pi}_n$ on $\mathcal{M}_c$ with dynamics f using the pessimistic value function $\underline{V}_r^{\hat{\pi}}(s_0)$ defined in Equation (8).

Lemma 3. *Suppose Assumptions 1 to 3 hold and the model $\mathcal{F}_n$ is well-calibrated according to Definition 1. Then the performance $J_r(\bar{\pi}_n, f)$ of the safe policy $\bar{\pi}_n$ induced by Algorithm 1 on $\mathcal{M}_c$ is lower-bounded by the performance $J_{\tilde{r}}(\pi_n, f)$ of the policy π_n on $\widetilde{\mathcal{M}}$ following the true dynamics f:*

$$
J_r(\bar{\pi}_n, f) \geq J_{\tilde{r}}(\pi_n, f).
$$

Proof. We consider for all timesteps k two cases:

1. *Never fall back to prior* (i.e. π_n is always safe). Then, by definition of $\widetilde{\mathcal{M}}$,

$$
J_{\tilde{r}}(\pi_n, f) = \mathbb{E}_{\pi_n, s_0} \Big[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \Big] = J_r(\bar{\pi}_n, f). \qquad (30)
$$

This equality follows by definition from $\widetilde{\mathcal{M}}$ and $\mathcal{M}_c$ being identical in case the agent never falls back to the safe prior $\hat{\pi}$ during the rollout and therefore $\pi_n = \bar{\pi}_n$.

2. *Agent falls back to safe prior* (i.e. at s_t action $\hat{a}_t = \hat{\pi}(s_t)$ is executed instead of potentially unsafe a_t). For the performance $J_r(\bar{\pi}_n, f)$ of $\bar{\pi}_n$ on $\mathcal{M}_c$ we get by combining the policy values of π_n and $\hat{\pi}$

$$J_r(\bar{\pi}_n, f) = \mathbb{E}_{\pi_n, s_0}\Big[\sum_{t=0}^{k-1} \gamma^t r(s_t, a_t)\Big] + \gamma^k\, V_r^{\hat{\pi}}(s_k), \tag{31}$$

whereas for the performance of $J_{\tilde{r}}(\pi_n, f)$ of π_n on $\widetilde{\mathcal{M}}$, we get

$$J_{\tilde{r}}(\pi_n, f) \overset{\text{(i)}}{=} \mathbb{E}_{\pi_n, s_0}\Big[\sum_{t=0}^{k-1} \gamma^t \tilde{r}(s_t, a_t) + \gamma^k \tilde{r}(s_k, \hat{a}_k) + \sum_{t=k+1}^{\infty} \gamma^t \tilde{r}(s_\dagger, \cdot)\Big] \tag{32}$$

$$\overset{\text{(ii)}}{=} \mathbb{E}_{\pi_n, s_0}\Big[\sum_{t=0}^{k-1} \gamma^t r(s_t, a_t) + \gamma^k\, \underline{V}_r^{\hat{\pi}}(s_k)\Big]. \tag{33}$$

Step (i) follows from the modified transitions in Equation (6) to the absorbing state $s_\dagger$ as soon as the agent would fall back to the safe prior $\hat{\pi}$, and (ii) follows from the terminal reward given in Equation (7) for transitioning to $s_\dagger$. Since $\underline{V}_r^{\hat{\pi}}$ is defined pessimistically in Equation (8), it holds that $\underline{V}_r^{\hat{\pi}}(s) \leq V_r^{\hat{\pi}}(s)$ and thereby follows

$$J_{\tilde{r}}(\pi_n, f) \leq \mathbb{E}_{\pi_n, s_0}\Big[\sum_{t=0}^{k-1} \gamma^t r(s_t, a_t)\Big] + \gamma^k\, V_r^{\hat{\pi}}(s_k) = J_r(\bar{\pi}_n, f), \tag{34}$$

which completes the proof.

□

Regret decomposition. Lemma 3 enables us to decompose $R(N)$ into two separate terms. The first term models the regret of executing $\bar{\pi}_n$ on $\mathcal{M}_c$, caused by falling back to the safe prior $\hat{\pi}$ when safely rolling out π_n with Algorithm 1. The second component determines the (constraint-free) regret of π_n on $\widetilde{\mathcal{M}}$.

Lemma 4. *Suppose Assumptions 1 to 4 hold and the model $\mathcal{F}_n$ is well-calibrated according to Definition 1. Let π_c^* be the optimal policy on $\mathcal{M}_c$ with $J_r(\pi_c^*, f) = \max_{\pi: J_c(\pi, f) < d} J_r(\pi, f)$ and applying Algorithm 1 to π_c^* yields $\bar{\pi}_{c,n}^*$ with $J_r(\bar{\pi}_{c,n}^*, f)$ for the current learned model $\mathcal{F}_n$ in episode n. Further, let π^* be the optimal policy on $\widetilde{\mathcal{M}}$ that has $J_{\tilde{r}}(\pi^*, f) = \max_\pi J_{\tilde{r}}(\pi, f)$ for Φ and $Q_{c,n}^{\hat{\pi}}$ in episode n. Given $J_{\tilde{r}}(\pi_n, f)$ denotes the performance of the behavioral policy π_n on $\widetilde{\mathcal{M}}$, the cumulative regret $R(N) = \sum_{n=1}^{N}(J_r(\pi_c^*, f) - J_r(\bar{\pi}_n, f))$ of $\bar{\pi}_n$ on $\mathcal{M}_c$ can be decomposed into:*

$$R(N) \leq \sum_{n=1}^{N} \Big(\underbrace{J_r(\pi_c^*, f) - J_r(\bar{\pi}_{c,n}^*, f)}_{\Delta_n^1 \,:\, \textit{P.E.R. of } \bar{\pi}_c^* \textit{ with } \Phi}\Big) + \sum_{n=1}^{N} \Big(\underbrace{J_{\tilde{r}}(\pi^*, f) - J_{\tilde{r}}(\pi_n, f)}_{\Delta_n^2 \,:\, \textit{P.E.R. of } \pi_n \textit{on } \tilde{\mathcal{M}}}\Big),$$

with the per-episode regret (P.E.R.) terms Δ_n^1 and Δ_n^2 related to the safe prior and the regret on $\widetilde{\mathcal{M}}$.

Proof. Starting with the cumulative regret definition in Equation (2) for $\mathcal{M}_c$, we upper bound $R(N)$ using

$$R(N) \overset{\text{(i)}}{=} \sum_{n=1}^{N}(J_r(\pi_c^*, f) - J_r(\bar{\pi}_n, f)) \leq \sum_{n=1}^{N}(J_r(\pi_c^*, f) - J_{\tilde{r}}(\pi_n, f)) \tag{35}$$

$$\overset{\text{(ii)}}{=} \sum_{n=1}^{N}(J_r(\pi_c^*, f) - J_r(\bar{\pi}_{c,n}^*, f) + J_r(\bar{\pi}_{c,n}^*, f) - J_{\tilde{r}}(\pi_n, f)), \tag{36}$$

where step (i) follows applying Lemma 3 and (ii) introducing a zero term by adding and subtracting $J_r(\bar{\pi}_{c,n}^*, f)$. Since $\mathcal{M}_c$ and $\widetilde{\mathcal{M}}$ are identical for safe policies (e.g. $\bar{\pi}_{c,n}^*$) by not transitioning to $s_\dagger$ (see Equations (6) and (7)) and π^* is the optimal policy on $\widetilde{\mathcal{M}}$ with $\mathcal{F}_n$ in episode n

$$J_r(\bar{\pi}_{c,n}^*, f) = J_{\tilde{r}}(\bar{\pi}_{c,n}^*, f) \leq J_{\tilde{r}}(\pi^*, f). \tag{37}$$

Thus, we obtain the following upper bound of $R(N)$ by plugging Equation (37) into Equation (36):

$$R(N) \leq \sum_{n=1}^{N} (J_r(\pi_c^*, f) - J_r(\bar{\pi}_{c,n}^*, f)) + \sum_{n=1}^{N} (J_{\tilde{r}}(\pi^*, f) - J_{\tilde{r}}(\pi_n, f)). \tag{38}$$

□

Δ_n^1 regret bound. Recall from Lemma 4 that the per episode regret of executing π_c^* with Φ for $\mathcal{F}_n$ in episode n on the planning MDP $\mathcal{M}_c$ is $\Delta_n^1 = J_r(\pi_c^*) - J_r(\bar{\pi}_{c,n}^*)$. We rewrite Δ_n^1 in terms of the distance between the optimal policy π_c^* and the safely rolled-out policy $\bar{\pi}_{c,n}^*$ in Lemma 5.

Lemma 5. *Suppose Assumptions 1 to 4 hold and the model $\mathcal{F}_n$ is calibrated according to Definition 1. Then the per-episode regret $\Delta_n^1 = J_r(\pi_c^*, f) - J_r(\bar{\pi}_{c,n}^*, f)$ admits the bound*

$$\Delta_n^1 \leq \mathbb{E}_{\bar{\pi}_c^*, s_0}\left[\left(L_r + \frac{\bar{R} L_f}{\sigma} \frac{\gamma}{1-\gamma} \right) \sum_{t=0}^{\infty} \gamma^t \, \mathbb{E}_{\substack{a_c^* \sim \pi_c^*(\cdot|s_t) \\ \bar{a}_{c,n}^* \sim \bar{\pi}_{c,n}^*(\cdot|s_t)}} \left[\|a_c^* - \bar{a}_{c,n}^*\| \right] \right],$$
$$s_{t+1} = f(s_t, a_t) + \omega_t, \quad a_t \sim \bar{\pi}_{c,n}^*(\cdot|s_t)$$

with the Lipschitz constants L_r, L_f for the reward and dynamics, and $\bar{R} = \max\{R_{max}, k_{max}\}$.

Proof. We bound the episodic regret with the advantage formulation [9, Lemma A.2]

$$\Delta_n^1 = \mathbb{E}_{\bar{\pi}_{c,n}^*, s_0}\left[\sum_{t=0}^{\infty} \gamma^t \, \mathbb{E}_{a_c^* \sim \pi_c^*(\cdot|s_t)} \left[A_r(\bar{\pi}_{c,n}^*, s_t, a_c^*) \right] \right], \tag{39}$$

with $A_{r,t}(\pi, s_t, a_t) = r(s_t, a_t) + \gamma \mathbb{E}_\omega [V_r^\pi(s_{t+1}) - V_r^\pi(s_t)]$ and $s_{t+1} = f(s_t, a_t) + \omega_t$.

$$\Delta_n^1 \overset{\text{(i)}}{=} \mathbb{E}_{\bar{\pi}_{c,n}^*, s_0}\left[\sum_{t=0}^{\infty} \gamma^t \, \mathbb{E}_{a_c^* \sim \pi_c^*(\cdot|s_t)} \left[\left(r(s_t, a_c^*) \right.\right.\right. \tag{40}$$
$$\left.\left.\left. + \gamma \left(\mathbb{E}_{s_{t+1}|s_t, \pi_c^*, \omega} \left[V_r^{\bar{\pi}_{c,n}^*}(s_{t+1}) \right] - V_r^{\bar{\pi}_{c,n}^*}(s_t) \right) \right) \right] \right]$$

$$\overset{\text{(ii)}}{=} \mathbb{E}_{\bar{\pi}_{c,n}^*, s_0}\left[\sum_{t=0}^{\infty} \gamma^t \, \mathbb{E}_{\substack{a_c^* \sim \pi^*(\cdot|s_t) \\ \bar{a}_{c,n}^* \sim \bar{\pi}_{c,n}^*(\cdot|s_t)}} \left[(r(s_t, a_c^*) - r(s_t, \bar{a}_{c,n}^*)) \right] \right] \tag{41}$$
$$+ \mathbb{E}_{\bar{\pi}_{c,n}^*, s_0}\left[\sum_{t=0}^{\infty} \gamma^{t+1} \left(\mathbb{E}_{s_{t+1}|s_t, \pi_c^*, \omega}[V_r^{\bar{\pi}_{c,n}^*}(s_{t+1})] - \mathbb{E}_{s_{t+1}|s_t, \bar{\pi}_{c,n}^*, \omega}[V_r^{\bar{\pi}_{c,n}^*}(s_{t+1})] \right) \right]$$

$$\overset{\text{(iii)}}{\leq} \mathbb{E}_{\bar{\pi}_{c,n}^*, s_0}\left[\sum_{t=0}^{\infty} \gamma^t \mathbb{E}_{\substack{a_c^* \sim \pi^*(\cdot|s_t) \\ \bar{a}_{c,n}^* \sim \bar{\pi}_{c,n}^*(\cdot|s_t)}} \left[\min\{L_r \|a_c^* - \bar{a}_{c,n}^*\|, R_{max}\} \right] \right. \tag{42}$$
$$+ \sum_{t=0}^{\infty} \gamma \sqrt{\max\{\mathbb{E}_{s_{t+1}|s_t, \pi_c^*, \omega_t}[R(s_{t+1})], \mathbb{E}_{s_{t+1}|s_t, \bar{\pi}_{c,n}^*, \omega_t}[R(s_{t+1})]\}}$$
$$\left. \times \gamma^t \mathbb{E}_{\substack{a_c^* \sim \pi^*(\cdot|s_t) \\ \bar{a}_{c,n}^* \sim \bar{\pi}_{c,n}^*(\cdot|s_t)}} \left[\min\left\{ \frac{L_f \|a_c^* - \bar{a}_{c,n}^*\|}{\sigma}, 1 \right\} \right] \right],$$

where step (i) follows from the advantage definition and (ii) uses the value function to derive the dependence on the reward difference in timestep t. We rewrite the advantage in terms of the expected action difference in step (iii) using the expectation difference under two Gaussians [53, lemma C.2] with $R(s) = \left(V_r^{\bar{\pi}_{c,n}^*}(s) \right)^2$. Step (iii) uses the Lipschitz constants L_r, L_f for the reward and the dynamics (Assumption 2). Further, we bound this by:

$$\Delta_n^1 \overset{\text{(i)}}{\leq} \mathbb{E}_{\bar{\pi}_{c,n}^*, s_0}\left[\sum_{t=0}^{\infty} \gamma^t \mathbb{E}_{\substack{a_c^* \sim \pi^*(\cdot|s_t) \\ \bar{a}_{c,n}^* \sim \bar{\pi}_{c,n}^*(\cdot|s_t)}} \left[\min\{L_r \|a_c^* - \bar{a}_{c,n}^*\|, R_{max}\} \right] \right. \tag{43}$$
$$\left. + \bar{R} \frac{\gamma}{1-\gamma} \sum_{t=0}^{\infty} \gamma^t \mathbb{E}_{\substack{a_c^* \sim \pi^*(\cdot|s_t) \\ \bar{a}_{c,n}^* \sim \bar{\pi}_{c,n}^*(\cdot|s_t)}} \left[\min\left\{ \frac{L_f \|a_c^* - \bar{a}_{c,n}^*\|}{\sigma}, 1 \right\} \right] \right]$$

$$\overset{\text{(ii)}}{\leq} \mathbb{E}_{\bar{\pi}^*_{c,n}, s_0}\left[\left(L_r + \frac{\bar{R} L_f}{\sigma} \frac{\gamma}{1-\gamma} \right) \sum_{t=0}^{\infty} \gamma^t \mathbb{E}_{\substack{a^*_c \sim \pi^*(\cdot|s_t) \\ \bar{a}^*_{c,n} \sim \bar{\pi}^*_{c,n}(\cdot|s_t)}} \left[\| a^*_c - \bar{a}^*_{c,n} \| \right] \right]. \tag{44}$$

Step (i) uses $\sqrt{\max\{\mathbb{E}_{\omega_t, s_{t+1}|s_t, \pi^*_c, \omega}[R(s_{t+1})], \mathbb{E}_{\omega_t, s_{t+1}|s_t, \bar{\pi}^*_{c,n}, \omega}[R(s_{t+1})]\}} \leq \bar{R}\frac{1}{1-\gamma}$ with $\bar{R} = \max\{R_{\text{max}}, k_{\text{max}}\}$ due to $r \geq 0$ and in (ii) we drop the $\min$ operator due to non-negativity. □

Lemma 5 expresses the regret in terms of the difference between actions taken by an optimal policy π^*_c and those taken by $\bar{\pi}^*_{c,n}$ in episode n with model $\mathcal{F}_n$. In the following, we use Lemma 5 to relate Δ^1_n to model uncertainties "collected" along trajectories induced by these policies. Bounding these uncertainties ensures that the learner improves its model of the dynamics after each episode.

Lemma 6. *Suppose Assumptions 1 to 4 hold and the model $\mathcal{F}_n$ is well calibrated according to Definition 1. Given Φ is defined as in Theorem 1, Δ^1_n depends on the accumulated and discounted uncertainty Σ_t along the trajectories of action a_t and the safe $\hat{\pi}$, and the diameter of the bounded action space $D_{\mathcal{A}}$. Given $\nu_n = (1 + \lambda_{pessimism})\lambda_{pessimism}$ with $\lambda_{pessimism}$ being defined as in Theorem 1,*

$$\Delta^1_n \leq \mathbb{E}_{\bar{\pi}^*_{c,n}, s_0}\left[\left(L_r + \frac{R_{max} L_f}{\sigma} \frac{\gamma}{1-\gamma} \right) \sum_{t=0}^{\infty} \gamma^t D_{\mathcal{A}} \min\left\{ \mathbb{E}_{a_t \sim \pi^*_c(\cdot|s_t), \omega_t}\left[\frac{\nu_n \Sigma_t(a_t, f)}{\delta_c} \right], 1 \right\} \right],$$

$$\Sigma_t(a_t, f) = \gamma^t \|\sigma_n(s_t, a_t)\| + \sum_{\tau=t+1}^{\infty} \mathbb{E}[\gamma^\tau_c \|\sigma_n(s_\tau, \hat{\pi}(s_\tau))\|], \quad s_{\tau+1} = f(s_\tau, \hat{\pi}(s_\tau)) + \omega_\tau$$

and $\delta_c = \min_t(\mathbb{E}_{a_t \sim \bar{\pi}^*_{c,n}(\cdot|s_t)}[d - (c_{<t} + \gamma^t c(s_t, a_t) + \gamma^{t+1}\mathbb{E}_{\omega_t}[V^{\hat{\pi}}_c(s_{t+1})])]) \in \mathbb{R}^+$.

Proof. We can upper bound the expected difference between the policies by the action bound $D_{\mathcal{A}}$ times the probability $\mathbb{P}$ of falling back to the safe prior

$$D_{\bar{\pi}^*_{c,n}}(s_t) = \mathbb{E}_{\substack{a^*_c \sim \pi^*(\cdot|s_t) \\ \bar{a}^*_{c,n} \sim \bar{\pi}^*_{c,n}(\cdot|s_t)}}\left[\| a^*_c - \bar{a}^*_{c,n} \| \right] \tag{45}$$

$$\overset{\text{(i)}}{\leq} D_{\mathcal{A}} \mathbb{P}_{a_t \sim \pi^*_c(\cdot|s_t)}\left(c_{<t} + \gamma^t Q^{\hat{\pi}}_{c,n}(s_t, a_t) \geq d \right) \tag{46}$$

$$\overset{\text{(ii)}}{=} D_{\mathcal{A}} \mathbb{P}_{a_t \sim \pi^*_c(\cdot|s_t)}\Big(c_{<t} + \gamma^t c(s_t, a_t) \tag{47}$$

$$+ \gamma^{t+1}\left(\mathbb{E}_{\omega_t}[\tilde{V}^{\hat{\pi}}_c(\mu_n(s'_t, a'_t) + \omega_t)] + \lambda_{\text{pessimism}} \Sigma_t(a_t, \mu_n) \right) \geq d \Big)$$

with $s'_{t+1} = \mu_n(s'_t, a'_t) + \omega_t$, $\Sigma_t(a_t, \mu_n) = \|\sigma_n(s_t, a_t)\| + \sum_{\tau=t+1}^{\infty} \mathbb{E}[\gamma^{\tau-t} \|\sigma_n(s'_\tau, \hat{\pi}(s'_\tau))\|]$ and $\tilde{V}^{\hat{\pi}}_c(s_0) = \sum_{t=0}^{\infty} \gamma^t c(s'_t, a'_t)$ following the model dynamics μ_n. Step (i) follows from the switching criterion in Theorem 1 and step (ii) by the definition in Equation (5). Next, we relate the uncertainty along trajectories under the model dynamics μ_n to those under the true dynamics f by

$$D_{\bar{\pi}^*_{c,n}}(s_t) \overset{\text{(i)}}{\leq} D_{\mathcal{A}} \mathbb{P}_{a_t \sim \pi^*_c(\cdot|s_t)}\Big(c_{<t} + \gamma^t c(s_t, a_t) + \gamma^{t+1}\Big(\mathbb{E}_{\omega_t}[\tilde{V}^{\hat{\pi}}_c(\mu_n(s'_t, a'_t) + \omega_t)] \tag{48}$$

$$+ \mathbb{E}_{\omega_t}[V^{\hat{\pi}}_c(f(s_t, a_t) + \omega_t)] - \mathbb{E}_{\omega_t}[V^{\hat{\pi}}_c(f(s_t, a_t) + \omega_t)]$$

$$+ \lambda_{\text{pessimism}}(\Sigma_t(a_t, \mu_n) + \Sigma_t(a_t, f) - \Sigma_t(a_t, f))\Big) \geq d \Big)$$

$$\overset{\text{(ii)}}{\leq} D_{\mathcal{A}} \mathbb{P}_{a_t \sim \pi^*_c(\cdot|s_t)}\Big(c_{<t} + \gamma^t c(s_t, a_t) + \gamma^{t+1}\Big(\mathbb{E}_{\omega_t}[V^{\hat{\pi}}_c(f(s_t, a_t) + \omega_t)] \tag{49}$$

$$+ (1 + \lambda_{\text{pessimism}})\lambda_{\text{pessimism}} \Sigma_t(a_t, f)\Big) \geq d \Big)$$

$$\overset{\text{(iii)}}{\leq} D_{\mathcal{A}} \mathbb{P}_{a_t \sim \pi^*_c(\cdot|s_t)}\left(c_{<t} + \gamma^t c(s_t, a_t) + \gamma^{t+1}\left(\mathbb{E}_{\omega_t}[V^{\hat{\pi}}_c(s_{t+1})] + \nu_n \Sigma_t(a_t, f) \right) \geq d \right) \tag{50}$$

We first expand the inequality in (i) with two zero terms. In step (ii), we bound the difference of the value and the uncertainty for the different dynamics using Sukhija et al. [76, Lemma A.5]. In (iii) we define $\nu_n = (1+\lambda_{\text{pessimism}})\lambda_{\text{pessimism}}$ and sample according to the true dynamics $s_{t+1} = f(s_t, a_t) + \omega_t$.

Consider the set of all policies that are safely reachable given $\hat{\pi}$

$$\Pi^{\hat{\pi}}_{<d} := \{\pi \mid \mathbb{E}_\pi[c_{<t} + \gamma^t c(s_t, a_t)] + \gamma^{t+1}\mathbb{E}_{\omega_t}[V_c^{\hat{\pi}}(s_{t+1})] < d, \forall t\}. \tag{51}$$

Since π_c^* is strictly feasible we have $\mathbb{E}_{\pi_c^*}[c_{<t} + \gamma^t c(s_t, a_t)] + \gamma^{t+1}\mathbb{E}_{\omega_t}[V_c^{\pi_c^*}(s_{t+1})] < d$; in addition, $V_c^{\hat{\pi}}(s) \leq V_c^{\pi_c^*}(s)$ for all initial states (cf. Assumption 4), therefore $\mathbb{E}_{\pi_c^*}[c_{<t} + \gamma^t c(s_t, a_t)] + \gamma^{t+1}\mathbb{E}_{\omega_t}[V_c^{\hat{\pi}}(s_{t+1})] < d$ implying that $\pi_c^* \in \Pi^{\hat{\pi}}_{<d}$. We define the smallest (time-wise) safety gap due to switching from π_c^* to $\hat{\pi}$ as

$$\delta_c := \min_t(\mathbb{E}_{a_t \sim \pi_c^*(\cdot|s_t)}[d - (c_{<t} + \gamma^t c(s_t, a_t) + \gamma^{t+1}\mathbb{E}_{\omega_t}[V_c^{\hat{\pi}}(s_{t+1})])]) \in \mathbb{R}_{>0}. \tag{52}$$

We thus rewrite the probability of invoking the safe prior $\hat{\pi}$

$$D_{\bar{\pi}^*_{c,n}}(s_t) = D_{\mathcal{A}}\mathbb{P}_{a_t \sim \pi_c^*(\cdot|s_t)}\Bigg(\nu_n \Sigma_t(a_t, f) \geq d - (c_{<t} + \gamma^t c(s_t, a_t) + \gamma^{t+1}\mathbb{E}_{\omega_t}[V_c^{\hat{\pi}}(s_{t+1})])\Bigg) \tag{53}$$

$$= D_{\mathcal{A}}\mathbb{P}_{a_t \sim \pi_c^*(\cdot|s_t)}\Bigg(\nu_n \Sigma_t(a_t, f) \geq \delta_c\Bigg). \tag{54}$$

Using *Markov's inequality*, we can bound the probability of invoking the prior by

$$D_{\bar{\pi}^*_{c,n}}(s_t) \leq D_{\mathcal{A}} \min\left\{\mathbb{E}_{a_t \sim \pi_c^*(\cdot|s_t), \omega_t}\left[\frac{\nu_n \Sigma_t(a_t, f)}{\delta_c}\right], 1\right\} \tag{55}$$

Finally, we insert Equation (55) into Lemma 5 and obtain

$$\Delta_n^1 \leq \mathbb{E}_{\bar{\pi}^*_{c,n}, s_0}\left[\left(L_r + \frac{\bar{R}L_f}{\sigma}\frac{\gamma}{1-\gamma}\right)\sum_{t=0}^{\infty}\gamma^t D_{\mathcal{A}} \min\left\{\mathbb{E}_{a_t \sim \pi_c^*(\cdot|s_t), \omega_t}\left[\frac{\nu_n \Sigma_t(a_t, f)}{\delta_c}\right], 1\right\}\right]. \tag{56}$$

□

Remark 1. Indeed, at the beginning of learning, π_c^* may not lie within the set $\{\pi \mid \mathbb{E}_\pi[c_{<t} + \gamma^t c(s_t, a_t)] + \gamma^{t+1}\mathbb{E}_{\omega_t}[\bar{V}_c^{\hat{\pi}}(s_{t+1})] < d, \forall t\} \subseteq \Pi^{\hat{\pi}}_{<d}$ as the pessimistic cost value $V_c^{\hat{\pi}} \leq \bar{V}_c^{\hat{\pi}}$ for all $s \in \mathcal{S}$ with probability $1 - \delta$ by definition (recall Equation (4)). Reducing pessimism such that $\bar{V}_c^{\hat{\pi}}$ converges to $V_c^{\hat{\pi}}$ effectively *expands* this set until it converges to $\Pi^{\hat{\pi}}_{<d}$. Moreover, let us consider the case where the prior policy does *not* satisfy the assumption that $V_c^{\hat{\pi}}(s) \leq V_c^{\pi_c^*}(s)$ for all $s \in \mathcal{S}$, i.e., the pessimistic prior policy $\hat{\pi}$ in fact accumulates *more* costs (in expectation) than an optimal policy π_c^* when executed on the true dynamics f. In this case, π_c^* is not guaranteed to lie within $\Pi^{\hat{\pi}}_{<d}$ and SOOPER converges to the optimum within the set of all feasible policies $\Pi^{\hat{\pi}}_{<d}$ that are safely reachable given $\hat{\pi}$.

Lemma 6 establishes an upper bound on the first regret term Δ_n^1, based on the probability of action $a_t \sim \pi_c^*(\cdot|s_t)$ to be unsafe given the model uncertainty. More concretely, even though actions $a_t \sim \pi_c^*(\cdot|s_t)$ in Equation (56) are determined by a *safe* policy π_c^*, they may be regarded unsafe due to limited knowledge during learning, quantified by epistemic model uncertainty. Since such trajectories would trigger $\hat{\pi}$ when executing Algorithm 1, we must relate the bound in Equation (56) to uncertainty induced along trajectories executed by the prior policy $\hat{\pi}$. Therefore, we relate in Lemma 7 these accumulated uncertainties when executing π_c^* freely in t, to trajectories induced under Algorithm 1, starting with the safe prior policy at timestep t. This relation is visualized in Figure 7.

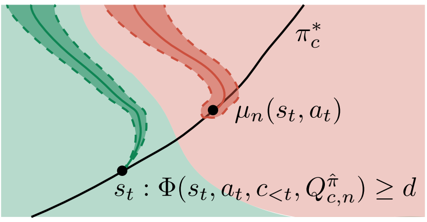


Figure 7: Relating the uncertainty of a safe trajectory (green) to a trajectory that executes π_c^* freely at t (i.e. not under Algorithm 1, in red) and therefore may be (possibly wrongly) considered unsafe due to model uncertainties.

Lemma 7. *Suppose Assumptions 1 to 4 hold and the model $\mathcal{F}_n$ is well-calibrated according to Definition 1, the accumulated uncertainty along a trajectory induced by executing π_c^* freely at t, quantified by $\mathbb{E}_{a_t\sim\pi_c^*(\cdot|s_t),\omega_t}[\Sigma_t(a_t, f)]$, can be upper-bounded by the uncertainty along a safe trajectory starting with the safe prior according to Algorithm 1:*

$$\mathbb{E}_{a_t\sim\pi_c^*(\cdot|s_t),\omega_t}[\Sigma_t(a_t, f)] \leq \left(1 + \frac{L_\sigma D_{\mathcal{A}}}{\sigma}\right)\sqrt{\frac{\gamma^t}{1-\gamma}}\mathbb{E}_{\omega_t}[\Sigma_t(\hat{\pi}(s_t), f)]. \tag{57}$$

Proof. Using the definition of Σ_t from Lemma 6

$$\mathbb{E}_{a_t\sim\pi_c^*(\cdot|s_t),\omega_t}[\Sigma_t(a_t, f)] \leq \left(1 + \frac{L_\sigma D_{\mathcal{A}}}{\sigma}\right)\sqrt{\frac{\gamma^t}{1-\gamma}}\sum_{\tau=t}^{\infty}\mathbb{E}_{\omega_t}[\sqrt{\gamma^\tau}\|\sigma_n(s_\tau, \hat{\pi}(s_\tau))\|], \tag{58}$$

we can bound the difference between the accumulated uncertainties of the two trajectories

$$\begin{aligned}\Delta\Sigma_t =& \mathbb{E}_{a_t\sim\pi_c^*(\cdot|s_t)}[\mathbb{E}_{\omega_t}[\Sigma_t(a_t, f)] - \mathbb{E}_{\omega_t}[\Sigma_t(\hat{\pi}(s_t), f)]] & (59)\\ \overset{\text{(i)}}{\leq}& \sqrt{\mathbb{E}_{\omega_t}\left[\left(\sum_{\tau=t+1}^{\infty}\mathbb{E}[\gamma^\tau\|\sigma_n(s_\tau, \hat{\pi}(s_\tau))\|]\right)^2\right]} & (60)\\ &\times\min\left\{\frac{\mathbb{E}_{a_t\sim\pi_c^*(\cdot|s_t)}[\|\sigma_n(s_t, a_t) - \sigma_n(s_t, \hat{\pi}(s_t))\|]}{\sigma}, 1\right\}\\ \overset{\text{(ii)}}{\leq}& \frac{L_\sigma D_{\mathcal{A}}}{\sigma}\sqrt{\mathbb{E}_{\omega_t}[\Sigma_t(\hat{\pi}(s_t), f)]^2}, & (61)\end{aligned}$$

by using the expectation difference under two Gaussians [53, Lemma C.2] in (i) and bound the uncertainty difference of under the different actions for the first timestep in (ii) using the closed set of actions with diameter $D_{\mathcal{A}}$ and the Lipschitz constant L_σ of the uncertainty. Thus we can further bound this term using $\mathbb{E}_{a_t\sim\pi_c^*(\cdot|s_t),\omega_t}[\Sigma_t] \leq \sqrt{\mathbb{E}_{\omega_t}[\Sigma_t(\hat{\pi}(s_t), f)]^2}$ and obtain

$$\begin{aligned}\mathbb{E}_{a_t\sim\pi_c^*(\cdot|s_t),\omega_t}[\Sigma_t(a_t, f)] &\leq \left(1 + \frac{L_\sigma D_{\mathcal{A}}}{\sigma}\right)\sqrt{\mathbb{E}_{\omega_t}[\Sigma_t(\hat{\pi}(s_t), f)^2]} & (62)\\ &\overset{\text{(i)}}{\leq} \left(1 + \frac{L_\sigma D_{\mathcal{A}}}{\sigma}\right)\sqrt{\mathbb{E}_{\omega_t}\left[\left(\sum_{\tau=t}^{\infty}\gamma^\tau\|\sigma_n(s_\tau, \hat{\pi}(s_\tau))\|\right)^2\right]} & (63)\\ &\overset{\text{(ii)}}{\leq} \left(1 + \frac{L_\sigma D_{\mathcal{A}}}{\sigma}\right)\sqrt{\mathbb{E}_{\omega_t}\left[\left(\sum_{\tau=t}^{\infty}\gamma^\tau\right)\left(\sum_{\tau=t}^{\infty}\gamma^\tau\|\sigma_n(s_\tau, \hat{\pi}(s_\tau))\|^2\right)\right]} & (64)\\ &\overset{\text{(iii)}}{\leq} \left(1 + \frac{L_\sigma D_{\mathcal{A}}}{\sigma}\right)\sqrt{\frac{\gamma^t}{1-\gamma}\left(\sum_{\tau=t}^{\infty}\mathbb{E}_{\omega_t}[\gamma^\tau\|\sigma_n(s_\tau, \hat{\pi}(s_\tau))\|^2]\right)} & (65)\\ &\overset{\text{(iv)}}{\leq} \left(1 + \frac{L_\sigma D_{\mathcal{A}}}{\sigma}\right)\sqrt{\frac{\gamma^t}{1-\gamma}\left(\sum_{\tau=t}^{\infty}\mathbb{E}_{\omega_t}[\sqrt{\gamma^\tau}\|\sigma_n(s_\tau, \hat{\pi}(s_\tau))\|]\right)^2} & (66)\\ &\leq \left(1 + \frac{L_\sigma D_{\mathcal{A}}}{\sigma}\right)\sqrt{\frac{\gamma^t}{1-\gamma}}\sum_{\tau=t}^{\infty}\mathbb{E}_{\omega_t}[\sqrt{\gamma^\tau}\|\sigma_n(s_\tau, \hat{\pi}(s_\tau))\|], & (67)\end{aligned}$$

where we get step (i) by restating the definition of $\Sigma_t(\hat{\pi}(s_t))$ from Lemma 6 and thereafter applying the *Cauchy-Schwarz* inequality in (ii). In step (iii), we factor out the infinite sum of γ^τ and can interchange the square with the summation in (iv) due to the non-negativity of $\|\sigma_n\|$. □

Following from the relation between uncertainty accumulated along trajectories that execute π_c^* freely at t and thereafter follow $\hat{\pi}$, and those that pessimistically invoke $\hat{\pi}$ in Lemma 7, we can bound the first regret term when episodes are collected safely using Algorithm 1.

Lemma 8. *Suppose Assumptions 1 to 4 hold and the model is calibrated according to Definition 1. Given the Lipschitz constants L_σ, with bounded actions $D_\mathcal{A}$, the first regret term Δ_n^1 is bounded with probability $1-\delta$ in terms of the safely rolled out policy $\bar{\pi}_{c,n}^*$:*

$$\Delta_n^1 \leq \mathbb{E}_{\bar{\pi}_c^*}\left[\eta_n \sum_{t=0}^{\infty} \mathbb{E}[\sqrt{\gamma^t} \|\sigma_n(s_t, a_t)\|]\right], \quad s_{t+1} = \mu_n(s_t, a_t) + \omega_t, \quad a_t \sim \bar{\pi}_{c,n}^*(\cdot|s_t),$$

where $\eta_n = \left(L_r + \frac{\bar{R}L_f}{\sigma}\frac{\gamma}{1-\gamma}\right)\frac{1}{(1-\gamma)^{\frac{3}{2}}} D_\mathcal{A}\left(1 + \frac{L_\sigma D_\mathcal{A}}{\sigma}\right)\frac{(1+\lambda_{\textit{pessimism}})^2 \lambda_{\textit{pessimism}}}{\delta_c}$.

Proof. Given the derived regret bound in Lemma 6, we bound the sum by the largest summation term during the safe rollout (i.e. executing π_c^* under Algorithm 1) using abbreviation $C' = \left(L_r + \frac{\bar{R}L_f}{\sigma}\frac{\gamma}{1-\gamma}\right)$.

$$\begin{aligned}
\Delta_n^1 \leq& \mathbb{E}_{\bar{\pi}_{c,n}^*, s_0}\left[C' \sum_{t=0}^{\infty} \gamma^t D_\mathcal{A} \min\left\{\mathbb{E}_{a_t \sim \pi_c^*(\cdot|s_t), \omega_t}\left[\frac{\nu_n \Sigma_t(a_t, f)}{\delta_c}\right], 1\right\}\right] && (68)\\
\leq& \mathbb{E}_{\bar{\pi}_{c,n}^*, s_0}\left[C' \frac{1}{1-\gamma} D_\mathcal{A} \max_t\left(\min\left\{\mathbb{E}_{a_t \sim \pi_c^*(\cdot|s_t), \omega_t}\left[\frac{\nu_n \Sigma_t(a_t, f)}{\delta_c}\right], 1\right\}\right)\right] && (69)\\
\leq& \mathbb{E}_{\bar{\pi}_{c,n}^*, s_0}\left[C' \frac{1}{1-\gamma} D_\mathcal{A} \min\left\{\max_t\left(\mathbb{E}_{a_t \sim \pi_c^*(\cdot|s_t), \omega_t}\left[\frac{\nu_n \Sigma_t(a_t, f)}{\delta_c}\right]\right), 1\right\}\right]. && (70)
\end{aligned}$$

As we invoke the prior for the maximal term we obtain together with Lemma 7

$$\begin{aligned}
\Delta_n^1 \leq& \mathbb{E}_{\bar{\pi}_{c,n}^*, s_0}\Bigg[C' \frac{1}{1-\gamma} D_\mathcal{A} \min\Bigg\{\max_t\Bigg(\left(1 + \frac{L_\sigma D_\mathcal{A}}{\sigma}\right)\sqrt{\frac{\gamma^t}{1-\gamma}} && (71)\\
&\times \frac{\nu_n}{\delta_c} \sum_{\tau=t}^{\infty} \mathbb{E}_{\omega_t}[\sqrt{\gamma^\tau} \|\sigma_n(s_\tau, \hat{\pi}(s_\tau))\|]\Bigg), 1\Bigg\}\Bigg]\\
\leq& \mathbb{E}_{\bar{\pi}_{c,n}^*, s_0}\left[C' \frac{1}{1-\gamma} D_\mathcal{A}\left(1 + \frac{L_\sigma D_\mathcal{A}}{\sigma}\right)\sqrt{\frac{1}{1-\gamma}}\frac{\nu_n}{\delta_c} \sum_{t=0}^{\infty} \mathbb{E}_{\omega_t}[\sqrt{\gamma^t} \|\sigma_n(s_t, a_t)\|]\right], && (72)\\
&s_{t+1} = f(s_t, a_t) + \omega_t, \quad a_t \sim \bar{\pi}_{c,n}^*(\cdot|s_t). && (73)
\end{aligned}$$

Since we now have the uncertainty along the entire safe rollout of $\bar{\pi}_{c,n}^*$ following the true dynamics f, we can now express the entire rollout in terms of the model dynamics μ_n, using [76, Lemma A.5]:

$$\begin{aligned}
\Delta_n^1 \overset{\text{(i)}}{\leq}& \mathbb{E}_{\bar{\pi}_{c,n}^*, s_0}\Bigg[C' \frac{1}{1-\gamma} D_\mathcal{A}\left(1 + \frac{L_\sigma D_\mathcal{A}}{\sigma}\right)\sqrt{\frac{1}{1-\gamma}}\frac{\nu_n}{\delta_c}\Bigg(\sum_{t=0}^{\infty} \mathbb{E}_{\omega_t}[\sqrt{\gamma^t} \|\sigma_n(s_t, a_t)\|] && (74)\\
&+ \sum_{t=0}^{\infty} \mathbb{E}_{\omega_t}[\sqrt{\gamma^t} \|\sigma_n(s_t', a_t')\|] - \sum_{t=0}^{\infty} \mathbb{E}_{\omega_t}[\sqrt{\gamma^t} \|\sigma_n(s_t', a_t')\|]\Bigg]\Bigg)\\
\overset{\text{(ii)}}{\leq}& \mathbb{E}_{\bar{\pi}_{c,n}^*, s_0}\Bigg[C' \frac{1}{1-\gamma} D_\mathcal{A}\left(1 + \frac{L_\sigma D_\mathcal{A}}{\sigma}\right)\sqrt{\frac{1}{1-\gamma}}\frac{\nu_n}{\delta_c} && (75)\\
&\times (1 + \lambda_{\text{pessimism}}) \sum_{t=0}^{\infty} \mathbb{E}_{\omega_t}[\sqrt{\gamma^t} \|\sigma_n(s_t', a_t')\|]\Bigg]\\
&s_{t+1} = f(s_t, a_t) + \omega_t, \quad s_{t+1}' = \mu_n(s_t', a_t') + \omega_t. && (76)
\end{aligned}$$

In step (i), we introduce a zero term by adding and subtracting the accumulated uncertainties along the model dynamics. Next, we bound in (ii) the difference of the uncertainties along the two different dynamics with [76, Lemma A.5]. Thus, the proof is complete by plugging back in C' and ν_n. □

Δ_n^2 regret bound. Let us now consider the second term Δ_n^2 that determines the regret of the policy π_n on $\widetilde{\mathcal{M}}$, given $\mathcal{F}_n$. Therefore, we first bound the value difference between the pessimistic value estimate $\underline{V}_r^{\hat{\pi}}$ and the true value of executing $\hat{\pi}$ on f. Leveraging this bound, we can upper-bound the value difference of executing π_n on $\mathcal{M}_c$ and $\widetilde{\mathcal{M}}$ in Lemma 10. Finally, we bound the second regret term Δ_n^2 in Lemma 11.

Lemma 9. *Suppose Assumptions 1 to 3 hold and the model $\mathcal{F}_n$ is well-calibrated according to Definition 1. Given the pessimistic cost value $\bar{V}_c^{\hat{\pi}}$ from Equation (4), we get with probability $1-\delta$*

$$\gamma^k |V_r^{\hat{\pi}}(s_k) - \underline{V}_r^{\hat{\pi}}(s_k)| \leq \lambda_n \sum_{t=k}^{\infty} \mathbb{E}\left[\gamma^t \|\sigma_n(\tilde{s}_t, \hat{\pi}(\tilde{s}_t))\|\right], \quad \lambda_n = \frac{\bar{R}\gamma}{1-\gamma} \frac{2(1+\sqrt{d_x})\beta_{n-1}(\delta)}{\sigma},$$

where $\bar{R} = \max\{R_{max}, k_{max}\}$. The trajectory can either be along the worst-case dynamics $\tilde{s}_{t+1} = \mu_n(\tilde{s}_t, \hat{\pi}(\tilde{s}_t)) + \alpha(1+\sqrt{d_x})\beta_{n-1}(\delta)\sigma_n(\tilde{s}_t, \hat{\pi}(\tilde{s}_t)) + \omega_t$, with $\alpha \in [-1, 1]$, initial state $\tilde{s}_k = s_k$, or the true dynamics $\tilde{s}_{t+1} = f(\tilde{s}_t, \hat{\pi}(\tilde{s}_t)) + \omega_t$, yielding the same bound.

Proof.

$$\gamma^k |V_r^{\hat{\pi}}(s_k) - \underline{V}_r^{\hat{\pi}}(s_k)| = \left|\mathbb{E}\left[\sum_{t=k}^{\infty} \gamma^t (V_r^{\hat{\pi}}(s_{t+1}) - V_r^{\hat{\pi}}(\tilde{s}_{t+1}))\right]\right| \tag{77}$$

$$\leq \sum_{t=k}^{\infty} \gamma^t \mathbb{E}[|\mathbb{E}_{\omega_t}(\underline{V}_r^{\hat{\pi}}(s_{t+1}) - \underline{V}_r^{\hat{\pi}}(\tilde{s}_{t+1}))|], \tag{78}$$

where $\tilde{s}_{t+1}$ is in each step the worst next state given $\mathcal{F}_n$, leading to the worst value (see Equation (4)). Thus with $\tilde{s}_{t+1} = \mu_n(\tilde{s}_t, \hat{\pi}(\tilde{s}_t)) + \alpha(1+\sqrt{d_x})\beta_{n-1}(\delta)\sigma_n(\tilde{s}_t, \hat{\pi}(\tilde{s}_t)) + \omega_t$, where $\alpha \in [-1, 1]$ and $\hat{R}(s) = (V_r^{\hat{\pi}})^2(s)$ it holds

$$\gamma^k |V_r^{\hat{\pi}}(s_k) - \underline{V}_r^{\hat{\pi}}(s_k)| \overset{\text{(i)}}{\leq} \sum_{t=k}^{\infty} \gamma \mathbb{E}\Bigg[\sqrt{\max\{\mathbb{E}_{\omega_t}[\hat{R}(s_{t+1})], \mathbb{E}_{\omega_t}[\hat{R}(\tilde{s}_{t+1})]\}} \tag{79}$$

$$\times \gamma^t \min\left\{\frac{\|f(s_t, \hat{\pi}(s_t)) - \mu_n(\tilde{s}_t, \hat{\pi}(\tilde{s}_t)) + \alpha(1+\sqrt{d_x})\beta_{n-1}(\delta)\sigma_n(\tilde{s}_t, \hat{\pi}(\tilde{s}_t))\|}{\sigma}, 1\right\}\Bigg]$$

$$\overset{\text{(ii)}}{\leq} \frac{\bar{R}\gamma}{1-\gamma} \frac{2(1+\sqrt{d_x})\beta_{n-1}(\delta)}{\sigma} \sum_{t=k}^{\infty} \mathbb{E}\left[\gamma^t \|\sigma_n(s_t, \hat{\pi}(s_t))\|\right]. \tag{80}$$

We rewrite the value difference dependent on the worst-case dynamics $\mu_n + \alpha(1+\sqrt{d_x})\beta_{n-1}(\delta)\sigma_n$ and the true dynamics f in step (i) using Kakade et al. [53, Lemma C.2], which can be expressed in terms of the model uncertainty in (ii) [84, Corollary 3]. □

Lemma 10. *Suppose Assumptions 1 to 3 hold and the model $\mathcal{F}_n$ is calibrated according to Definition 1. We consider the following definitions*

$$J_{\tilde{r}}(\pi_n, f) = \mathbb{E}_{\pi, s_0}\left[\sum_{t=0}^{\infty} \gamma^t \tilde{r}(s_t, a_t)\right],$$

$$a_t \sim \pi_n(\cdot|s_t), \quad s_{t+1} = \begin{cases} f(s_t, a_t) + \omega_t & \Phi(s_t, a_t, c_{<t}, Q_{c,n}^{\hat{\pi}}) \geq d, s_t \neq s_\dagger, \\ s_\dagger & \textit{otherwise,} \end{cases}$$

$$J_{\tilde{r}}(\pi_n, \mu_n) = \mathbb{E}_{\pi, s_0}\left[\sum_{t=0}^{\infty} \gamma^t \tilde{r}(\tilde{s}_t, \tilde{a}_t)\right],$$

$$\tilde{a}_t \sim \pi_n(\tilde{s}_t), \quad \tilde{s}_{t+1} = \begin{cases} \mu_n(\tilde{s}_t, \tilde{a}_t) + \omega_t & \Phi(\tilde{s}_t, a_t, c_{<t}, Q_{c,n}^{\hat{\pi}}) \geq d, \tilde{s}_t \neq s_\dagger, \\ s_\dagger & \textit{otherwise,} \end{cases}$$

$$\lambda_n = \frac{\bar{R}\gamma}{1-\gamma} \frac{2(1+\sqrt{d_x})\beta_{n-1}(\delta)}{\sigma}.$$

Then we have for all $n \geq 0$, with probability $1-\delta$

$$|J_{\tilde{r}}(\pi_n, \mu_n) - J_{\tilde{r}}(\pi_n, f)| \leq \lambda_n \sum_{t=0}^{\infty} \mathbb{E}_{\bar{\pi}_n, s_0}\left[\gamma^t \|\sigma_{n-1}(\tilde{s}_t, \tilde{a}_t)\|\right]$$

$$|J_{\tilde{r}}(\pi_n, \mu_n) - J_{\tilde{r}}(\pi_n, f)| \leq \lambda_n \sum_{t=0}^{\infty} \mathbb{E}_{\bar{\pi}_n, s_0}\left[\gamma^t \|\sigma_{n-1}(s_t, a_t)\|\right].$$

Proof. We prove the first inequality $|J_{\tilde{r}}(\pi_n,\mu_n) - J_{\tilde{r}}(\pi_n, f)| \leq \lambda_n \sum_{t=0}^{\infty} \mathbb{E}_{\pi_n,s_0}\left[\gamma^t \|\sigma_{n-1}(\tilde{s}_t,\tilde{a}_t)\|\right]$ and the second one holds for a analogous argument. We extend the argument of Sukhija et al. [76, Lemma A.5] to account for the termination state $s_\dagger$.

$$|J_{\tilde{r}}(\pi_n,\mu_n) - J_{\tilde{r}}(\pi_n, f)| = \left|\mathbb{E}_{\pi_n,s_0,\mu_n}\left[\sum_{t=0}^{\infty}\gamma^{t+1}(V^{\pi_n}_{r,f}(\tilde{s}_{t+1}) - V^{\pi_n}_{r,f}(s_{t+1}))\right]\right|, \tag{81}$$

where $V^{\pi}_{r,f}$ is the value following π under dynamics f, and we define the states $\tilde{s}_{t+1}$ and s_{t+1} as

$$s_{t+1} = \begin{cases} f(\tilde{s}_t,\tilde{a}_t) + \omega_t & \Phi(\tilde{s}_t,\tilde{a}_t,c_{<t},Q^{\hat{\pi}}_{c,n}) \geq d, \tilde{s}_t \neq s_\dagger, \\ s_\dagger & \text{otherwise}, \end{cases}, \quad \tilde{a}_t \sim \pi_n(\cdot|\tilde{s}_t), \tag{82}$$

$$\tilde{s}_{t+1} = \begin{cases} \mu_n(\tilde{s}_t,\tilde{a}_t) + \omega_t & \Phi(\tilde{s}_t,\tilde{a}_t,c_{<t},Q^{\hat{\pi}}_{c,n}) \geq d, \tilde{s}_t \neq s_\dagger, \\ s_\dagger & \text{otherwise}. \end{cases}, \quad \tilde{a}_t \sim \pi_n(\cdot|\tilde{s}_t). \tag{83}$$

Let $\frac{(1+\sqrt{d_s})\beta_n(\delta)}{\sigma}R(s) = \left(V^{\pi}_{r,f}(s)\right)^2$ and with $R(s) \leq \lambda_n$, if π_n does not invoke the safe prior $\hat{\pi}$, we obtain the following bound following Sukhija et al. [76, Lemma A.5]

$$\begin{aligned}
|J_{\tilde{r}}(\pi_n,\mu_n) - J_{\tilde{r}}(\pi_n, f)| \leq & \sum_{t=0}^{\infty}\gamma\mathbb{E}_{\pi_n,s_0}\Bigg[\sqrt{\max\{\mathbb{E}_{\omega_t}[R(\tilde{s}_{t+1})],\mathbb{E}_{\omega_t}[R(s_{t+1})]\}} & (84)\\
& \times\gamma^t\min\left\{\frac{\|s_{t+1}-\tilde{s}_{t+1}\|}{\sigma},1\right\}\Bigg]. \\
\leq & \sum_{t=0}^{\infty}\gamma\mathbb{E}_{\pi_n,s_0}\Bigg[\sqrt{\max\{\mathbb{E}_{\omega_t}[R(\tilde{s}_{t+1})],\mathbb{E}_{\omega_t}[R(s_{t+1})]\}} & (85)\\
& \times\gamma^t\min\left\{\frac{\|f(\tilde{s}_t,\tilde{a}_t)-\mu_n(\tilde{s}_t,\tilde{a}_t)\|}{\sigma},1\right\}\Bigg] \\
\leq & \frac{\bar{R}\gamma}{1-\gamma}\frac{(1+\sqrt{d_x})\beta_{n-1}(\delta)}{\sigma}\sum_{t=0}^{\infty}\mathbb{E}_{\pi_n,s_0}\left[\gamma^t\|\sigma_n(\tilde{s}_t,\tilde{a}_t)\|\right]. & (86)
\end{aligned}$$

In case the agent invokes $\hat{\pi}$ at some timestep $t = k$, we formulate the regret as

$$\begin{aligned}
|J_{\tilde{r}}(\pi_n,\mu_n) - J_{\tilde{r}}(\pi_n, f)| \overset{\text{(i)}}{\leq} & \sum_{t=0}^{k-1}\gamma\mathbb{E}_{\pi_n,s_0}\Bigg[\sqrt{\max\{\mathbb{E}_{\omega_t}[R(\tilde{s}_{t+1})],\mathbb{E}_{\omega_t}[R(s_{t+1})]\}} & (87)\\
& \times\gamma^t\min\left\{\frac{\|s_{t+1}-\tilde{s}_{t+1}\|}{\sigma},1\right\}\Bigg] + \gamma^k|V^{\hat{\pi}}_r(\tilde{s}_k) - \underline{V}^{\hat{\pi}}_r(\tilde{s}_k)| \\
& + \sum_{t=k+1}^{\infty}\sqrt{\mathbb{E}_{\omega_t}[R(s_\dagger)]}\gamma^t\min\left\{\frac{\|s_\dagger - s_\dagger\|}{\sigma},1\right\} \\
\overset{\text{(ii)}}{=} & \sum_{t=0}^{k-1}\gamma\mathbb{E}_{\pi_n,s_0}\Bigg[\sqrt{\max\{\mathbb{E}_{\omega_t}[R(\tilde{s}_{t+1})],\mathbb{E}_{\omega_t}[R(s_{t+1})]\}} & (88)\\
& \times\gamma^t\min\left\{\frac{\|s_{t+1}-\tilde{s}_{t+1}\|}{\sigma},1\right\}\Bigg] + \gamma^k|V^{\hat{\pi}}_r(\tilde{s}_k) - \underline{V}^{\hat{\pi}}_r(\tilde{s}_k)|, \\
\overset{\text{(iii)}}{\leq} & \sum_{t=0}^{k-1}\gamma\mathbb{E}_{\pi_n,s_0}\Bigg[\sqrt{\max\{\mathbb{E}_{\omega_t}[R(\tilde{s}_{t+1})],\mathbb{E}_{\omega_t}[R(s_{t+1})]\}} & (89)\\
& \times\gamma^t\min\left\{\frac{\|f(\tilde{s}_t,\pi(\tilde{s}_t))-\mu_n(\tilde{s}_t,\pi(\tilde{s}_t))\|}{\sigma},1\right\}\Bigg] \\
& + \frac{\bar{R}\gamma}{1-\gamma}\frac{2(1+\sqrt{d_x})\beta_{n-1}(\delta)}{\sigma}\sum_{t=k}^{\infty}\mathbb{E}_{\hat{\pi},\tilde{s}_k}\left[\gamma^t\|\sigma_n(\tilde{s}_t,\hat{\pi}(\tilde{s}_t))\|\right] \\
\overset{\text{(iv)}}{\leq} & \frac{\bar{R}\gamma}{1-\gamma}\frac{(1+\sqrt{d_x})\beta_{n-1}(\delta)}{\sigma}\sum_{t=0}^{k-1}\mathbb{E}_{\pi_n,s_0}[\gamma^t|\sigma_n(\tilde{s}_t,\tilde{a}_t)\|] & (90)
\end{aligned}$$

$$+ \frac{\bar{R}\gamma}{1-\gamma}\frac{2(1+\sqrt{d_x})\beta_{n-1}(\delta)}{\sigma}\sum_{t=k}^{\infty}\mathbb{E}_{\hat{\pi},\tilde{s}_k}\left[\gamma^t\|\sigma_n(\tilde{s}_t,\hat{\pi}(\tilde{s}_t))\|\right]$$

$$\overset{\text{(v)}}{\leq} \frac{\bar{R}\gamma}{1-\gamma}\frac{2(1+\sqrt{d_x})\beta_{n-1}(\delta)}{\sigma}\sum_{t=0}^{\infty}\mathbb{E}_{\bar{\pi}_n,s_0}\left[\gamma^t\|\sigma_n(\tilde{s}_t,\tilde{a}_t)\|\right]. \tag{91}$$

In step (i), we split the infinite sum along the trajectory, when the policy prior $\hat{\pi}$ is invoked. We therefore follow the trajectory along the model dynamics in Equation (83). Further, we eliminate the difference term for the terminal state $s_\dagger$ in (ii). In step (iii), we use Lemma 9 to bound the value difference between the true value $V_r^{\hat{\pi}}$ along f and the pessimistic value $\underline{V}_r^{\hat{\pi}}$. Using Sukhija et al. [84, Corollary 3] we bound the difference between the next states in (iv). Finally, in step (v), we formulate the bound over $\bar{\pi}_n$, allowing us to upper bound both terms with the uncertainty using the safely rolled out policy $\bar{\pi}_n$ and abbreviate $\lambda_n = \frac{\bar{R}\gamma}{1-\gamma}\frac{2(1+\sqrt{d_x})\beta_{n-1}(\delta)}{\sigma}$. □

Lemma 11. *Suppose Assumptions 1 to 3 hold and the model is well-calibrated according to definition 1. Given the optimal policy π^*, then we obtain the following per-episode regret on $\widetilde{\mathcal{M}}$ in episode n, $\Delta_n^2 = J_{\tilde{r}}(\pi^*, f) - J_{\tilde{r}}(\pi, f)$ with $\lambda_n = \frac{\bar{R}\gamma}{1-\gamma}\frac{2(1+\sqrt{d_x})\beta_{n-1}(\delta)}{\sigma}$ with probability $(1-\delta)$*

$$\Delta_n^2 = J_{\tilde{r}}(\pi^*, f) - J_{\tilde{r}}(\pi_n, f) \leq J_{\tilde{r}}(\pi^*, \mu_n) + \lambda_n \sum_{t=0}^{\infty}\mathbb{E}_{\bar{\pi}^*,s_0}\left[\gamma^t\|\sigma_n(s_t,a_t)\|\right] - J_{\tilde{r}}(\pi_n, f),$$

where $s_{t+1} = \mu_n(s_t,a_t) + \omega_t$

Proof. We can show this by applying Lemma 10 to an optimal policy π^* on $\widetilde{\mathcal{M}}$. □

Intrinsic rewards for exploration and expansion. Next, we show with Lemmas 12 and 13 that the policy π_n in Equation (9) achieves both optimism and expansion by upper-bounding the sum of both regret terms $\Delta_n = \Delta_n^1 + \Delta_n^2$.

Lemma 12. *Given Assumptions 1 to 4 hold and the model $\mathcal{F}_n$ is well-calibrated according to Definition 1, the policy π_n as defined in Equation (9)*

$$\begin{aligned}\pi_n =& \arg\max_{\pi} \mathbb{E}_{\tilde{p},\pi,s_0}\left[\sum_{t=0}^{\infty}\Big(\gamma^t\tilde{r}(s_t,a_t) + (\gamma^t\lambda_{\textit{explore}} + \sqrt{\gamma^t}\lambda_{\textit{expand}})\|\sigma_n(s_t,a_t)\|\Big)\right],\\ & s_{t+1} = \mu_n(s_t,a_t) + \omega_t, \quad a_t \sim \pi(\cdot|s_t),\end{aligned}$$

with $\lambda_{\textit{explore}} = 3\lambda_n$ (Lemma 11) and $\lambda_{\textit{expand}} = 3\eta_n$ (Lemma 8), satisfies with probability $1-\delta$:

$$\begin{aligned}\Delta_n =& \Delta_n^1 + \Delta_n^2 \leq \eta_n \sum_{t=0}^{\infty}\mathbb{E}_{\pi_c^*,s_0}[\sqrt{\gamma^t}\|\sigma_n(s_t,a_t)\|] + J_{\tilde{r}}(\pi^*, \mu_n)\\ & + \lambda_n \sum_{t=0}^{\infty}\mathbb{E}_{\bar{\pi}^*,s_0}\left[\gamma^t\|\sigma_n(s_t,a_t)\|\right] - J_{\tilde{r}}(\pi_n, f)\\ \leq & \sum_{t=0}^{\infty}\mathbb{E}_{\pi_n,s_0}[(\gamma^t\lambda_{\textit{explore}} + \sqrt{\gamma^t}\lambda_{\textit{expand}})|\sigma_n(s_t,a_t)\|] + J_{\tilde{r}}(\pi_n,\mu_n) - J_{\tilde{r}}(\pi_n, f).\end{aligned}$$

Proof. Recall that by construction π_n maximizes

$$J(\pi) = \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty}\gamma^t\,\tilde{r}(s_t,a_t)) + \sum_{t=0}^{\infty}\Big(\sqrt{\gamma^t}3\eta_n + \gamma^t 3\lambda_n\Big)\|\sigma_n(s_t,a_t))\|\right]. \tag{92}$$

Hence

$$J(\pi_n) \geq J(\pi) \quad \forall\pi. \tag{93}$$

Apply Equation (93) first to the policy $\bar{\pi}_{c,n}^*$ (i.e. the policy induced by using π_c^* and then invoking the safe prior policy $\hat{\pi}$). By the linearity of expectations and non-negativity of all terms,

$$J(\pi_n) \geq J_{\tilde{r}}(\bar{\pi}_{c,n}^*, \mu_n) + \mathbb{E}_{\bar{\pi}_{c,n}^*,s_0}\left[3\eta_n\sum_{t=0}^{\infty}\sqrt{\gamma^t}\|\sigma_n(s_t,a_t)\| + 3\lambda_n\sum_{t=0}^{\infty}\gamma^t\|\sigma_n(s_t,a_t)\|\right]. \tag{94}$$

Next apply Equation (93) to the optimal policy π^* on $\mathcal{F}_n$, yielding

$$J(\pi_n) \geq J_{\tilde{r}}(\pi^*, \mu_n) + \mathbb{E}_{\pi^*, s_0}\left[3\eta_n \sum_{t=0}^{\infty} \sqrt{\gamma^t} \|\sigma_n(s_t, a_t)\| + 3\lambda_n \sum_{t=0}^{\infty} \gamma^t \|\sigma_n(s_t, a_t))\|\right]. \tag{95}$$

Adding Equation (94) and Equation (95) results in

$$\begin{aligned} 2J(\pi_n) \geq & J_{\tilde{r}}(\bar{\pi}^*_{c,n}, \mu_n) + \mathbb{E}_{\bar{\pi}^*_{c,n}, s_0}\left[3\eta_n \sum_{t=0}^{\infty} \sqrt{\gamma^t} \|\sigma_n(s_t, a_t)\| + 3\lambda_n \sum_{t=0}^{\infty} \gamma^t \|\sigma_n(s_t, a_t)\|\right] \\ & + J_{\tilde{r}}(\pi^*, \mu_n) + \mathbb{E}_{\pi^*, s_0}\left[3\eta_n \sum_{t=0}^{\infty} \sqrt{\gamma^t} \|\sigma_n(s_t, a_t)\| + 3\lambda_n \sum_{t=0}^{\infty} \gamma^t \|\sigma_n(s_t, a_t))\|\right]. \end{aligned} \tag{96}$$

Let us derive a bound for $J_{\tilde{r}}(\bar{\pi}^*_{c,n}, \mu_n)$, that we can relate to the optimal policy π^* for $\widetilde{\mathcal{M}}$.

$$\begin{aligned} J_{\tilde{r}}(\pi^*, \mu_n) & \overset{\text{(i)}}{\leq} J_{\tilde{r}}(\pi^*, f) + \lambda_n \mathbb{E}_{\pi^*, s_0}\left[\sum_{t=0}^{\infty} \gamma^t \|\sigma_n(s_t, a_t)\|\right] && (97) \\ & \overset{\text{(ii)}}{\leq} J_r(\bar{\pi}^*, f) + \lambda_n \mathbb{E}_{\bar{\pi}^*, s_0}\left[\sum_{t=0}^{\infty} \gamma^t \|\sigma_n(s_t, a_t)\|\right] && (98) \\ & \overset{\text{(iii)}}{\leq} J_r(\pi^*_c, f) + \lambda_n \mathbb{E}_{\bar{\pi}^*, s_0}\left[\sum_{t=0}^{\infty} \gamma^t \|\sigma_n(s_t, a_t)\|\right] && (99) \\ & \overset{\text{(iv)}}{\leq} J_r(\bar{\pi}^*_{c,n}, f) + \lambda_n \mathbb{E}_{\bar{\pi}^*, s_0}\left[\sum_{t=0}^{\infty} \gamma^t \|\sigma_n(s_t, a_t)\|\right] && (100) \\ & \quad + \eta_n \mathbb{E}_{\bar{\pi}^*_{c,n}, s_0}\left[\sum_{t=0}^{\infty} \sqrt{\gamma^t} \|\sigma_n(s_t, a_t)\|\right]. \end{aligned}$$

Step (i) follows from Lemma 10 and (ii) is derived from the definition of $\widetilde{\mathcal{M}}$ in Equations (6) and (7). Next, step (iii) uses the optimality of π^*_c and finally step (iv) uses the upper bound of the first regret term Δ^1_n in Lemma 8. Hence, we can lower bound parts of the term by $J_{\tilde{r}}(\pi^*, \mu_n)$. Therefore, $2J(\pi_n)$ in Equation (96) is lower bounded by

$$\begin{aligned} 2J(\pi_n) \geq & 2J_{\tilde{r}}(\pi^*, \mu_n) + 2\eta_n \mathbb{E}_{\bar{\pi}^*_{c,n}, s_0}\left[\sum_{t=0}^{\infty} \sqrt{\gamma^t} \|\sigma_n(s_t, a_t)\|\right] \\ & + 3\lambda_n \mathbb{E}_{\bar{\pi}^*_{c,n}, s_0}\left[\sum_{t=0}^{\infty} \gamma^t \|\sigma_n(s_t, a_t)\|\right] + 3\eta_n \mathbb{E}_{\bar{\pi}^*, s_0}\left[\sum_{t=0}^{\infty} \sqrt{\gamma^t} \|\sigma_n(s_t, a_t)\|\right] \\ & + \mathbb{E}_{\bar{\pi}^*, s_0}\left[2\lambda_n \sum_{t=0}^{\infty} \gamma^t \|\sigma_n(s_t, a_t)\|\right]. \end{aligned} \tag{101}$$

Since all terms are non-negative, we can drop the terms $3(\lambda_n + \eta_n)\mathbb{E}_{\bar{\pi}^*_{c,n}, s_0}\left[\sum_{t=0}^{\infty} \gamma^t \|\sigma_n(s_t, a_t)\|\right]$ and obtain by dividing both sides by two:

$$\begin{aligned} 2J(\pi_n) \geq & 2J_{\tilde{r}}(\pi^*, \mu_n) + 2\lambda_n \mathbb{E}_{\bar{\pi}^*, s_0}\left[\sum_{t=0}^{\infty} \gamma^t \|\sigma_n(s_t, a_t)\|\right] && (102) \\ & + \mathbb{E}_{\bar{\pi}^*_{c,n}, s_0}\left[2\eta_n \sum_{t=0}^{\infty} \sqrt{\gamma^t} \|\sigma_n(s_t, a_t)\|\right] \\ J(\pi_n) \geq & J_{\tilde{r}}(\pi^*, \mu_n) + \lambda_n \mathbb{E}_{\bar{\pi}^*, s_0}\left[\sum_{t=0}^{\infty} \gamma^t \|\sigma_n(s_t, a_t)\|\right] && (103) \\ & + \mathbb{E}_{\bar{\pi}^*_{c,n}, s_0}\left[\eta_n \sum_{t=0}^{\infty} \sqrt{\gamma^t} \|\sigma_n(s_t, a_t)\|\right]. \end{aligned}$$

By definition of $J(\pi_n)$ and Equation (103), subtracting $J_{\tilde{r}}(\pi_n, f)$ completes the proof. □

In Lemma 12 we show that the objective upper bounds the performance and the per-episode regret on the *model dynamics*. Next, we show in Lemma 13 that we can also bound the regret on the *true dynamics* f using $\bar{\pi}_n$.

Lemma 13. *Given Assumptions 1 to 4 hold and the model $\mathcal{F}_n$ is well-calibrated according to Definition 1, the per-episode regret in n, $\Delta_n = \Delta_n^1 + \Delta_n^2$, is upper-bounded with probability of at least $1-\delta$ for all $n > 0$ by*

$$\Delta_n \leq (3\lambda_n^2 + 4\lambda_n + 3\eta_n^2 + 3\eta_n)) \sum_{t=0}^{\infty} \mathbb{E}_{\bar{\pi}_n, s_0}[\sqrt{\gamma^t} \|\sigma_n(s_t, a_t)\|],$$

for $s_t = f(s_t, a_t) + \omega_t$ following the true dynamics with $a_t \sim \bar{\pi}_n(\cdot | s_t, c_{<t}, Q^{\hat{\pi}}_{c,n})$.

Proof. Using the derivation in Lemma 12, we can upper bound the regret in terms of π_n by

$$\Delta_n \overset{\text{(i)}}{\leq} \sum_{t=0}^{\infty} \mathbb{E}_{\pi_n, s_0}[(\gamma^t \lambda_{\text{explore}} + \sqrt{\gamma^t} \lambda_{\text{expand}}) \|\sigma_n(s'_t, a'_t)\|] + J_{\tilde{r}}(\pi_n, \mu_n) - J_{\tilde{r}}(\pi_n, f) \tag{104}$$

$$\overset{\text{(ii)}}{\leq} 3(\eta_n + \lambda_n) \mathbb{E}_{\bar{\pi}_n, s_0}\left[\sum_{t=0}^{\infty} \sqrt{\gamma^t} \|\sigma_n(s'_t, a'_t)\|\right] + J_{\tilde{r}}(\pi_n, \mu_n) - J_{\tilde{r}}(\pi_n, f), \tag{105}$$

where $s'_t = \mu_n(s'_t, a'_t) + \omega_t$ follows the true dynamics with $a'_t \sim \pi_n(\cdot | s'_t)$ and we insert Lemma 12 in (i). Step (ii) follows from $\gamma \in (0, 1)$ and the uncertainty along the safely executed policy $\bar{\pi}_n$ being larger than for π_n, which terminates in $s_\dagger$. We next use Lemma 10 to bound $J_{\tilde{r}}(\pi_n, \mu_n) - J_{\tilde{r}}(\pi_n, f)$ and obtain

$$\Delta_n \leq 3(\eta_n + \lambda_n) \mathbb{E}_{\bar{\pi}_n, s_0}\left[\sum_{t=0}^{\infty} \sqrt{\gamma^t} \|\sigma_n(s'_t, a'_t)\|\right] + \lambda_n \mathbb{E}_{\bar{\pi}_n, s_0}\left[\sum_{t=0}^{\infty} \sqrt{\gamma^t} \|\sigma_n(s_t, a_t)\|\right] \tag{106}$$

$$\begin{aligned}
&\overset{\text{(i)}}{=} 3\eta_n \mathbb{E}_{\bar{\pi}_n, s_0}\left[\sum_{t=0}^{\infty} \sqrt{\gamma^t} \|\sigma_n(s_t, a_t)\|\right] + 3\eta_n \left(\mathbb{E}_{\bar{\pi}_n, s_0}\left[\sum_{t=0}^{\infty} \sqrt{\gamma^t} \|\sigma_n(s'_t, a'_t)\|\right]\right. \\
&\quad \left. - \mathbb{E}_{\bar{\pi}_n, s_0}\left[\sum_{t=0}^{\infty} \sqrt{\gamma^t} \|\sigma_n(s_t, a_t)\|\right]\right) + 3\lambda_n \left(\mathbb{E}_{\bar{\pi}_n, s_0}\left[\sum_{t=0}^{\infty} \sqrt{\gamma^t} \|\sigma_n(s'_t, a'_t)\|\right]\right. \\
&\quad \left. - \mathbb{E}_{\bar{\pi}_n, s_0}\left[\sum_{t=0}^{\infty} \sqrt{\gamma^t} \|\sigma_n(s_t, a_t)\|\right]\right) + 4\lambda_n \mathbb{E}_{\bar{\pi}_n, s_0}\left[\sum_{t=0}^{\infty} \sqrt{\gamma^t} \|\sigma_n(s_t, a_t)\|\right],
\end{aligned} \tag{107}$$

where step (i) introduces zero terms by adding and subtracting accumulated uncertainties. We further bound the per episode regret using non-negativity and Lemma 10

$$\Delta_n \leq (3\lambda_n^2 + 4\lambda_n + 3\eta_n^2 + 3\eta_n) \mathbb{E}_{\bar{\pi}_n, s_0}\left[\sum_{t=0}^{\infty} \sqrt{\gamma^t} \|\sigma_n(s_t, a_t)\|\right]. \tag{108}$$

□

In Lemma 13, we derived an upper bound on the per episode regret Δ_n, based on the policy $\bar{\pi}_n$ executed on the true dynamics $s_{t+1} = f(s_t, a_t) + \omega_t$. Using Chowdhury and Gopalan [85], we obtain a sublinear cumulative regret for our safe exploration algorithm SOOPER.

Theorem 2 (Sublinear cumulative regret)**.** *Suppose Assumptions 1 to 4 hold and the model $\mathcal{F}_n$ is well calibrated according to Definition 1. Then, Algorithm 2 guarantees with probability $1-\delta$*

$$R(N) = \sum_{n=1}^{N} \left(J_r(\pi_c^*, f) - J_r(\bar{\pi}_n, f)\right) \leq \mathcal{O}\left(\Gamma^{7/2}_{N\log(N)} \sqrt{N}\right),$$
$$\text{and } J_c(\bar{\pi}_n, f) \leq d, \; \forall n \in \{1, \ldots, N\},$$

that is, the cumulative regret grows sub-linearly in N, while satisfying the constraint throughout learning by Theorem 1.

Proof. Given the per episode regret bound in Lemma 13, we sum over all episodes $n = 1, \dots, N$

$$R(N) = \sum_{n=1}^{N} \Delta_n^1 + \Delta_n^2 \leq \sum_{n=1}^{N} (3\lambda_n^2 + 4\lambda_n + 3\eta_n^2 + 3\eta_n) \sum_{t=0}^{\infty} \mathbb{E}_{\bar{\pi}_n, s_0} \left[\sqrt{\gamma^t} \|\sigma_n(s_t, a_t)\| \right]. \tag{109}$$

As by definition [85], for the well-calibrated model $\mathcal{F}_n$, $\beta_n \leq \beta_N$ for all $n = 1, \dots, N$, we rewrite

$$\begin{aligned} R(N) &\leq (4\lambda_N + 3(\lambda_N^2 + \eta_N^2 + \eta_N)) \sum_{n=1}^{N} \sum_{t=0}^{\infty} \mathbb{E}_{\bar{\pi}_n, s_0} \left[\sqrt{\gamma^t} \|\sigma_n(s_t, a_t)\| \right] & (110) \\ &\overset{\text{(i)}}{\leq} ((4\lambda_N + 3(\lambda_N^2 + \eta_N^2 + \eta_N)) \sqrt{N} \sqrt{\sum_{n=1}^{N} \mathbb{E}_{s_0} \left[\left(\sum_{t=0}^{\infty} \mathbb{E}_{\bar{\pi}_n} \left[\sqrt{\gamma^t} \|\sigma_n(s_t, a_t)\| \right] \right)^2 \right]} & (111) \\ &\overset{\text{(ii)}}{\leq} (4\lambda_N + 3(\lambda_N^2 + \eta_N^2 + \eta_N)) \sqrt{\frac{N}{1 - \sqrt{\gamma}}} \sqrt{\sum_{n=1}^{N} \mathbb{E}_{s_0} \left[\sum_{t=0}^{\infty} \mathbb{E}_{\bar{\pi}_n} \left[\sqrt{\gamma^t} \|\sigma_n(s_t, a_t)\|^2 \right] \right]}. & (112) \end{aligned}$$

We use *Cauchy-Schwarz* in step (i) for the outer sum over episodes $n = 1, \dots, N$ and in (ii) for the inner sum over timesteps t. Further, we can bound the regret term for episodes $n = 1, \dots, N$, using the truncated horizon $T_n = -\frac{\log(n)}{\log(\sqrt{\gamma})}$, by decomposing the term into

$$\begin{aligned} &\sum_{n=1}^{N} \mathbb{E}_{s_0} \left[\sum_{t=0}^{\infty} \mathbb{E}_{\bar{\pi}_n} \left[\sqrt{\gamma^t} \|\sigma_n(s_t, a_t)\|^2 \right] \right] & (113) \\ &= \sum_{n=1}^{N} \left(\sum_{t=0}^{T_n - 1} \mathbb{E}_{\bar{\pi}_n} \left[\sqrt{\gamma^t} \|\sigma_n(s_t, a_t)\|^2 \right] + \sum_{t=T_n}^{\infty} \mathbb{E}_{\bar{\pi}_n} \left[\sqrt{\gamma^t} \|\sigma_n(s_t, a_t)\|^2 \right] \right) & (114) \\ &\leq \sum_{n=1}^{N} \sum_{t=0}^{T_n - 1} \mathbb{E}_{\bar{\pi}_n} \left[\sqrt{\gamma^t} \|\sigma_n(s_t, a_t)\|^2 \right] + \sum_{t=1}^{N} \gamma^{\frac{T_n}{2}} \frac{k_{max}^2}{1 - \gamma} & (115) \\ &\overset{\text{(i)}}{\leq} \sum_{n=1}^{N} \sum_{t=0}^{T_n - 1} \mathbb{E}_{\bar{\pi}_n} \left[\sqrt{\gamma^t} \|\sigma_n(s_t, a_t)\|^2 \right] + \sum_{n=1}^{N} \frac{1}{n} \frac{k_{max}^2}{1 - \gamma} & (116) \\ &\overset{\text{(ii)}}{\leq} \sum_{n=1}^{N} \sum_{t=0}^{T_n - 1} \mathbb{E}_{\bar{\pi}_n} \left[\sqrt{\gamma^t} \|\sigma_n(s_t, a_t)\|^2 \right] + \frac{k_{max}^2}{1 - \gamma} (\log(N) + 1). & (117) \end{aligned}$$

Step (i) follows from $\frac{T_n}{2} = -\frac{\log(n)}{\log(\gamma)}$, with $\frac{\log(n)}{\log(\gamma)} = \log_\gamma(n)$ and hence $\gamma^{-\log_\gamma(n)} = n^{-1}$. In (ii), we use the fact that $\frac{1}{n}$ is non-decreasing in n and therefore due to Lehman et al. [86, Theorem 3.9.1] we bound $\sum_{n=1}^{N} \frac{1}{n} \leq \log(N) + 1$. Further, we bound the first term $\sum_{n=1}^{N} \sum_{t=0}^{T_n - 1} \mathbb{E}_{\bar{\pi}_n} \left[\sqrt{\gamma^t} \|\sigma_n(s_t, a_t)\|^2 \right]$ as in Sukhija et al. [76, theorem 5.7]:

$$\leq (4\lambda_N + 3(\lambda_N^2 + \eta_N^2 + \eta_N)) \sqrt{\frac{R_\gamma N \Gamma_{N \log(N)}}{1 - \sqrt{\gamma}} + \frac{R_\gamma \sigma_{max}^2 N (\log(N) + 1)}{1 - \gamma}}, \tag{118}$$

where $R_\gamma = \frac{s_{max}}{\log(1 + s_{max})}$, with $s_{max} = \frac{\sigma^{-2} d_x \sigma_{max}^2}{1 - \sqrt{\gamma}}$. As T_n is non-decreasing, we fix T_N and since $\lambda_N \propto \frac{\beta_N}{1 - \gamma}$, $\eta_n \propto \frac{\beta_N^3}{1 - \gamma}$, we obtain from the definition of β_N from Chowdhury and Gopalan [85] that

$$R_N \leq \mathcal{O}\left(\Gamma_{N \log(N)}^{7/2} \sqrt{N} \right). \tag{119}$$

□

C Theoretical Extension to World Models

For our theoretical analysis in Appendices A and B, we assume, for simplicity in presentation, that the reward and cost functions are known. All arguments can be extended to the more general setting of unknown costs and rewards, which are learned as part of a "world model", in addition to the unknown dynamics f. In practice, we consider this general setting in our experiments (Section 5). Below we briefly illustrate how one can extend the analysis above to this more general case.

First, one needs to impose an additional RKHS assumption on the cost and reward functions, analogous to Assumption 3 for the unknown dynamics f. Accordingly, we jointly learn with the unknown dynamics f a calibrated model for the reward and cost function with nominal predictions μ_n^r, μ_n^c and epistemic uncertainty σ_n^r, σ_n^c. Hence, it holds $\forall (s,a) \in \mathcal{S} \times \mathcal{A} : r(s,a) \in \{r' \mid |r' - \mu_n^r| \leq \beta_n \sigma_n^r\}$ and $\forall (s,a) \in \mathcal{S} \times \mathcal{A} : c(s,a) \in \{c' \mid |c' - \mu_n^c| \leq \beta_n \sigma_n^c\}$.

Given only access to the calibrated model predictions, we modify the reward and cost structure on the planning MDP $\widetilde{\mathcal{M}}$ as follows:

$$\tilde{r}(s_t, a_t) \coloneqq \begin{cases} \underline{V}_r^{\hat{\pi}}(s_t) & \text{if } \Phi(a_t, s_t, c_{<t}, Q_{c,n}^{\hat{\pi}}) \geq d, \\ 0 & \text{if } s_t = s_\dagger, \\ \mu_n^r(s_t, a_t) + \beta_n \sigma_n^r(s_t, a_t) & \text{otherwise,} \end{cases} \tag{120}$$

$$\tilde{c}(s_t, a_t) \coloneqq \mu_n^c(s_t, a_t) + \beta_n \sigma_n^c(s_t, a_t). \tag{121}$$

By utilizing the above cost and reward, we demonstrate safe exploration and achieve sublinear cumulative regret, even in the presence of unknown cost and reward functions.

Safety guarantee for unknown costs. Under the modified cost $\tilde{c}$ in Equation (121), the update of the pessimistic Q-value is given by:

$$Q_{\tilde{c},n}^{\hat{\pi}}(s,a) \coloneqq \mathbb{E}_{\hat{\pi}}\left[\sum_{t=0}^{\infty} \gamma^t (\tilde{c}(s_t, a_t) + \lambda_{\text{pessimism}} \|\sigma_n(s_t, a_t)\|) \;\middle|\; s_0 = s, a_0 = a\right]. \tag{122}$$

For this modified update using $\tilde{c}$, it is possible to show that Lemma 2 holds using uncertain predictions. One can derive it by bounding the difference between the value $\tilde{V}_c^{\hat{\pi}}(s)$ following the model dynamics with the true cost function, and the value $\tilde{V}_{\tilde{c}}^{\hat{\pi}}(s)$ following the model dynamics with the mean predictions

$$\begin{aligned} |\tilde{V}_c^{\hat{\pi}}(s) - \tilde{V}_{\tilde{c}}^{\hat{\pi}}(s)| &= \left|\mathbb{E}_{\hat{\pi}}\left[\sum_{t=0}^{\infty} \gamma^t (c(s_t, a_t) - \mu_n^c(s_t, a_t))\right] \mid s_0 = s\right| & (123) \\ &\leq \mathbb{E}_{\hat{\pi}}\left[\sum_{t=0}^{\infty} \gamma^t \left|c(s_t, a_t) - \mu_n^c(s_t, a_t)\right| \mid s_0 = s\right] & (124) \\ &\leq \mathbb{E}_{\hat{\pi}}\left[\sum_{t=0}^{\infty} \gamma^t \beta_n \sigma_n^c(s_t, a_t) \mid s_0 = s\right], & (125) \\ s_{t+1} &= \mu_n(s_t, a_t) + \omega_t, \quad s_0 = s. & (126) \end{aligned}$$

Hence, by bringing the actually observable value $\tilde{V}_{\tilde{c}}^{\hat{\pi}}(s)$ to the right side, it follows that

$$\begin{aligned} \tilde{V}_c^{\hat{\pi}}(s) &\leq \tilde{V}_{\tilde{c}}^{\hat{\pi}}(s) + \mathbb{E}_{\hat{\pi}}\left[\sum_{t=0}^{\infty} \gamma^t \beta_n \sigma_n^c(s_t, a_t) \mid s_0 = s\right] & (127) \\ &= \mathbb{E}_{\hat{\pi}}\left[\sum_{t=0}^{\infty} \gamma^t (\mu_n^c(s_t, a_t) + \beta_n \sigma_n^c(s_t, a_t)) \mid s_0 = s\right] & (128) \\ &= \mathbb{E}_{\hat{\pi}}\left[\sum_{t=0}^{\infty} \gamma^t \tilde{c}(s_t, a_t) \mid s_0 = s\right], & (129) \\ s_{t+1} &= \mu_n(s_t, a_t) + \omega_t, \quad a_t = \hat{\pi}(s_t). & (130) \end{aligned}$$

Using the derived upper bound and plugging it into Equation (20) yields

$$\begin{aligned} \bar{V}_c^{\hat{\pi}}(s) &\leq \mathbb{E}_{\hat{\pi}}\left[\sum_{t=0}^{\infty} \gamma^t \left(\tilde{c}(s_t, a_t) + \lambda_{\text{pessimism}} \|\sigma_n(s_t, a_t)\|\right) \mid s_0 = s\right], & (131) \\ s_{t+1} &= \mu_n(s_t, a_t) + \omega_t, \quad a_t = \hat{\pi}(s_t). & (132) \end{aligned}$$

Hence, the proof of Theorem 1 holds using $Q^{\hat{\pi}}_{\tilde{c},n}(s,a)$ from Equation (122) in Equation (29). We can also upper-bound this in a more compact form by assuming the same kernel k for the cost and the dynamics prediction, so that we have $\beta_n \sigma^c_n(s_t,a_t) + \lambda_{\text{pessimism}} \|\sigma_n(s_t,a_t)\| \leq (\lambda_{\text{pessimism}} + \beta_n)\|\sigma^{f,c}_n(s_t,a_t)\|$. This follows from the non-negativity of σ and with $\sigma^{f,c}_n$ being defined as the concatenation of the prediction uncertainties for the dynamics σ_n and the cost σ^c_n.

Regret bound for unknown reward and cost function. Similar to the modified cost update in Equation (122), we use $\tilde{r}$ in Equation (120), and optimize for

$$\pi_n = \arg\max_{\pi} \mathbb{E}_{\tilde{p},\pi,s_0}\left[\sum_{t=0}^{\infty}\left(\gamma^t \tilde{r}(s_t,a_t) + (\gamma^t \lambda_{\text{explore}} + \sqrt{\gamma^t}\lambda_{\text{expand}})\|\sigma_n(s_t,a_t)\|\right)\right], \tag{133}$$

$$s_{t+1} = \mu_n(s_t,a_t) + \omega_t, \quad a_t \sim \pi(\cdot|s_t). \tag{134}$$

For the derivation of the regret bound, the first regret term is affected by enlarging the pessimistic additive cost with $\lambda_{\text{pessimism}} + \beta_n$. This contributes an additive β_n term, which does not change the proportionality of $\eta_n \propto \frac{\beta^3_n}{1-\gamma}$. For the second regret term, we obtain an additional term that quantifies the uncertainty in the reward prediction, similar to the derivation in Equation (129). Hence, we obtain the additive term in Lemma 11 while following the mean dynamics and using the model's mean predictions μ^r_n.

$$\Delta^2_n = J_{\tilde{r}}(\pi^*, f) - J_{\tilde{r}}(\pi_n, f) \tag{135}$$

$$\leq J_{\tilde{r}}(\pi^*, \mu_n) + \lambda_n \sum_{t=0}^{\infty} \mathbb{E}_{\bar{\pi}^*,s_0}\left[\gamma^t \|\sigma_n(s_t,a_t)\|\right] - J_{\tilde{r}}(\pi_n, f) \tag{136}$$

$$\leq \sum_{t=0}^{\infty} \mathbb{E}_{\bar{\pi}^*}\left[\gamma^t(\mu^r_n(s_t,a_t) + \beta_n \sigma^r_n(s_t,a_t)) \mid s_0 = s\right] \tag{137}$$

$$+ \lambda_n \sum_{t=0}^{\infty} \mathbb{E}_{\bar{\pi}^*,s_0}\left[\gamma^t \|\sigma_n(s_t,a_t)\|\right] - J_{\tilde{r}}(\pi_n, f). \tag{138}$$

By definition of the modified objective in Equation (133), we still upper bound in Lemma 12 the regret with the behavioral policy π_n, so that the following proof steps still hold. Further, we can again assume that the dynamics and the reward are part of the same kernel, resulting in the bound $\beta_n \sigma^r_n(s_t,a_t) + \lambda_n \|\sigma_n(s_t,a_t)\| \leq (\lambda_n + \beta_n)\|\sigma^{f,r}_n(s_t,a_t)\|$, where $\sigma^{f,r}_n$ is defined as the concatenation of the prediction uncertainties of the dynamics σ_n and the reward σ^r_n. Thereby, also for the second regret term, the proportionality remains $\lambda_n \propto \frac{\beta_n}{1-\gamma}$. Consequently, the regret bound remains unchanged $R_N \leq \mathcal{O}\left(\Gamma^{7/2}_{N\log(N)}\sqrt{N}\right)$ for unknown reward and cost functions.

D Additional Experiments

We conduct further empirical analysis of SOOPER, focusing on the following key aspects: **(i)** demonstrating the role of prior policies in accelerating learning; **(ii)** leveraging terminations for safe policy improvement; **(iii)** ablating the role of Algorithm 1; **(iv)** robustness of SOOPER to prior policies with varying pessimism levels; **(v)** the necessity of acting pessimistically for safety, and **(vi)** the influence of optimism on the agent's regret. The goal of these ablations is to provide additional empirical evidence to support the different design choices of SOOPER.

Safe transfer under task misspecification. We demonstrate the effectiveness of policy priors compared to "backup" policies that are commonly considered in previous works. For this, we reduce the size of the goal in the PointGoal1 task from SafetyGym (see Figure 16) from 0.3 to 0.15. This makes learning harder, as additional exploration is required to find the goal. We compare the following methods: **(i)** SOOPER that uses a policy trained on the original 0.3 goal task as a safe prior; **(ii)** SAILR that uses a hand-crafted "backup" policy that brakes the robot upon triggering, following SAILR's "advantage-based filtering". The policy π_n in this case is initialized with the same policy prior that SOOPER uses; **(iii)** SOOPER that uses the policy prior only within Algorithm 1, but learns its policy π_n completely from scratch; **(iv)** finally, a baseline that uses the hand-crafted braking backup policy, falling back to it as done in in SAILR, while training the policy π_n from scratch. Figure 8 demonstrates two results: **(a)** as expected, both baselines that learn from scratch fail to converge within the training budget of 200 episodes. **(b)** On the other hand, when finetuning the prior policy, using a hand-crafted braking "backup" policy indeed satisfies the constraint, however at the cost of slower learning compared to SOOPER. This result illustrates the importance of using policy priors not only to maintain safety, but also to guide and accelerate learning.

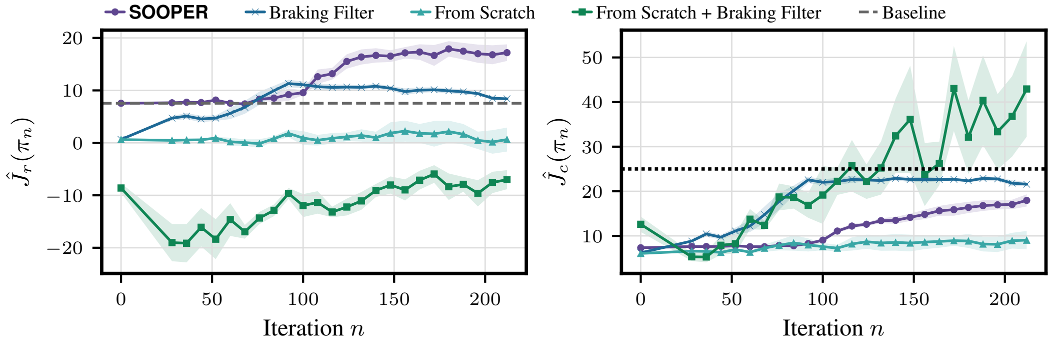


Figure 8: Learning curves of the objective and constraint for PointGoal1 goal size transfer. SOOPER performs significantly better using the prior policy trained on a larger goal size, significantly helping in exploration when learning to reach a smaller goal.

Learning from terminations. We study the effect of learning from terminations on the agent's ability to reduce its dependency on the policy prior. To this end, we ablate the use of terminations in Equations (6) and (7) and compare SOOPER with a variant that does not terminate in $\widetilde{\mathcal{M}}$. In Figure 9, we observe a decrease in the incurred costs for SOOPER with terminations, while SOOPER without terminations still encounters high costs for both CartpoleSwingup and RaceCar. If the optimal unconstrained policy violates the safety budget, the agent trained without terminations fails to achieve the optimal safe performance. This failure occurs because, in the absence of termination signals, the agent receives no feedback about triggering the prior policy during online rollouts.

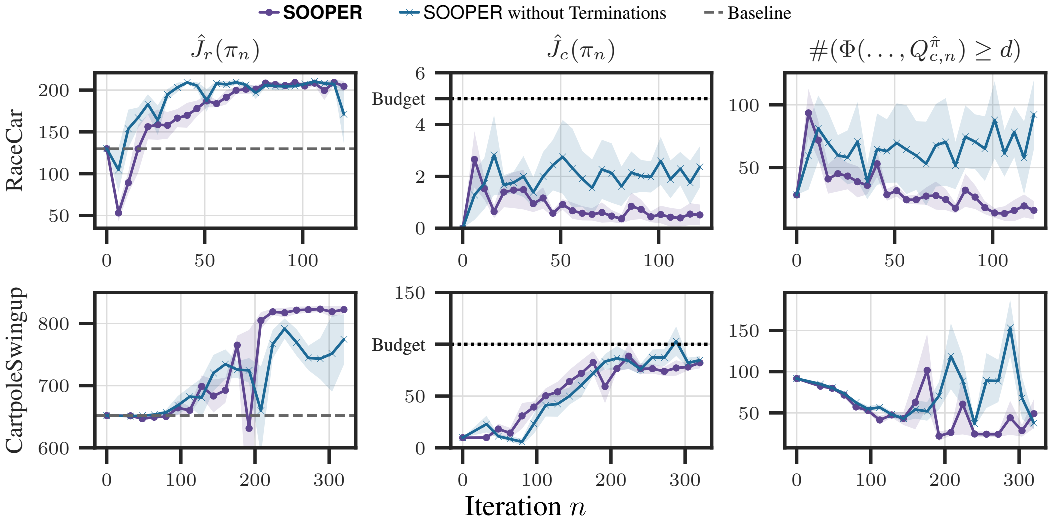


Figure 9: Learning curves of the objective, constraint, and number of times the safe prior is triggered during online rollouts. The number of times SOOPER uses the prior shrinks over training.

Robustness to policy priors. Different degrees of sim-to-real gap or lack of data coverage in the offline setting require increasing levels of conservatism. We demonstrate that SOOPER can effectively learn from policy priors with varying levels of conservatism and initial performance. For this, we train three prior policies on PointGoal2 for different sim-to-real gaps with sufficient pessimism levels and show that SOOPER converges to optimal performance while satisfying the constraints using all three policy priors in Figure 10. These results highlight SOOPER's ability to safely explore in highly uncertain environments, given conservative policy priors.

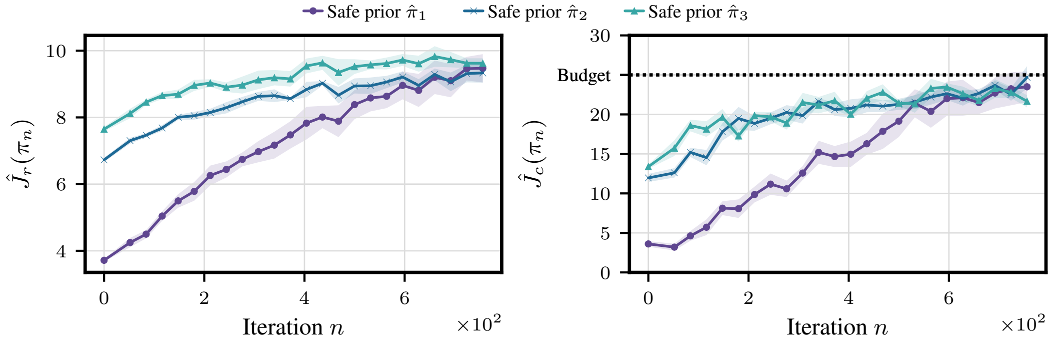


Figure 10: Learning curves of the objective and the constraint, starting from different safe policy priors with various levels of conservatism, differing in initial performances.

Pessimism for safe exploration. Given the previous ablation on conservatism of safe policy priors, we now study the scale of pessimism used *during* learning. As shown for CartpoleSwingup in Figure 11, using increased $\lambda_{\text{pessimism}}$, the agent consistently satisfies the constraints without significant sacrifice in performance. Moreover, SOOPER demonstrates robustness to increased $\lambda_{\text{pessimism}}$, and showcases higher performance for lower chosen pessimism values. Crucially, this behavior is expected as it reflects an inherent tradeoff between pessimism and exploration.

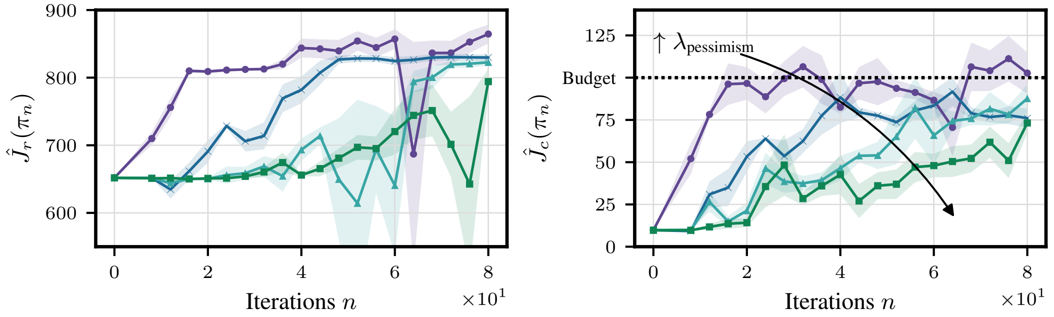


Figure 11: Learning curves of the objective and the constraint for SOOPER with increasing $\lambda_{\text{pessimism}}$. Pessimism improves robustness of constraint satisfaction.

Optimistic planning for sublinear regret. We now study the influence of SOOPER's intrinsic reward on its performance, safety during learning and on the cumulative regret. Specifically, we study how varying λ_{expand} together with λ_{explore} improves SOOPER's performance in the CartpoleSwingup task. Intuitively, high values of $\lambda_{\text{expand}} + \lambda_{\text{explore}}$ may lead to excessive exploration, while small values of $\lambda_{\text{expand}} + \lambda_{\text{explore}}$ may not explore enough. As visible in Figure 12, empirically, the cumulative regret $\frac{1}{N}R(N)$ exhibits a single optimum, since for small optimism scales the agent under-explores and fails to expand the safe region, and for large optimism scale the agent over-explores, which reduces short-term greediness and lowers cumulative reward over the finite horizon. As N increases, short-term suboptimality due to over-exploration decreases and the cumulative reward becomes less sensitive to the optimism scale, consistent with the sublinear-regret guarantee for sub-linear growth in N.

Safe offline-to-online on real hardware. We repeat our hardware experiment in Section 5 on the race car however now use SOOPER using an offline-trained policy. We collect 25K real-world transitions, corresponding to roughly ten minutes of real-time data. To collect this data, we use the simulation-trained policy used in our sim-to-real experiment as baseline. This dataset is then used to train a pessimistic policy using MOPO [62] with a primal-dual solver for the constraint. We introduce pessimism by penalizing the reward and cost with model uncertainty [see 62]. We repeat our hardware experiment with five random seeds, reporting the mean and standard error of the accumulated rewards and costs of each iteration n in Figure 13. As shown, SOOPER leverages the offline-trained policy prior to learn near-optimal policy while satisfying the constraint throughout learning.

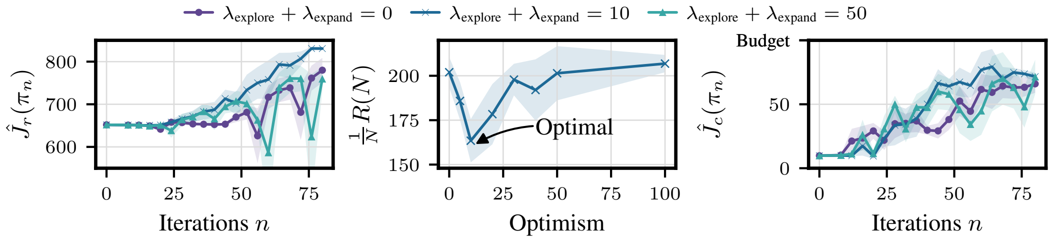


Figure 12: Learning curves of the objective and the constraint, as well as cumulative regret, for different optimism scales. A good value for the optimism value can be identified, where both under- and over-exploration are avoided and regret is minimized.

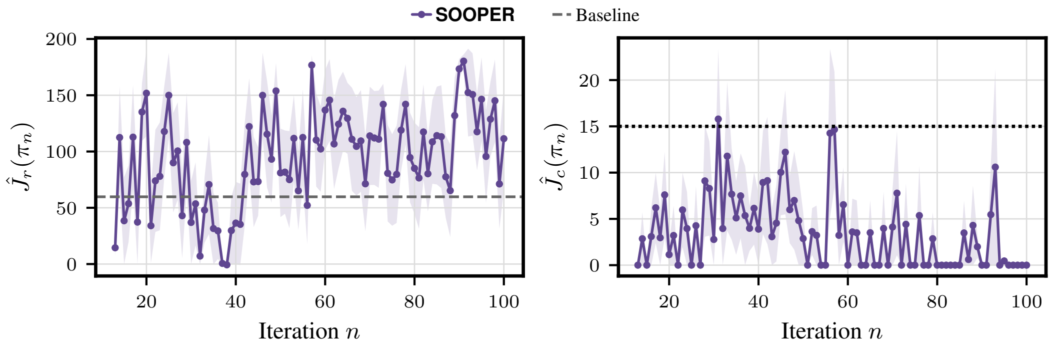


Figure 13: Performance and safety on the real race car given a prior policy that is trained via a fixed offline data. SOOPER improves the performance while satisfying the constraint during learning.

Learning directly from costs. While we demonstrate that SOOPER can learn safety from an unconstrained MDP with termination, our method can be extended to learning directly from costs via model-based rollouts with any off-the-shelf CMDP solver. While this variant enjoys the safety guarantee of Theorem 1 using Algorithm 1, it is not guaranteed to learn an optimal policy. We instantiate this variant with CRPO and refer to this baseline as SafeCRPO. The evaluations across various tasks show that SOOPER's Algorithm 1 can be paired with alternative CMDP solvers and ensures safe learning. However, SafeCRPO consistently learns slower and achieves lower performance than SOOPER. This demonstrates the advantage of SOOPER using terminations paired with Algorithm 1 to learn safely and guarantee convergence.

Additional learning curves. Figure 15 complements Figure 3 in Section 5 and shows the learning curves on all sim-to-sim tasks. SOOPER consistently achieves near-optimal performance without violating the constraint during learning. In addition to the baselines presented in Section 5, we include another CMDP solver that is based on Log Barriers [46]. This solver is theoretically guaranteed to maintain feasibility during learning by penalizing iterates that approach the boundary of the feasible set. For all of our sim-to-sim experiments, we train the safe policy priors following SPiDR [74]. SPiDR leverages domain randomization to introduce the conservatism required for constraint satisfaction when deploying RL policies under distribution shifts. In particular, to obtain pessimistic policies, SPiDR penalizes the cost according to the disagreement in next-state predictions of the randomized dynamics.

E SafetyGym

PointGoal environments. In these tasks, the agent must navigate to a target location while avoiding hazards which include free-moving vases and designated hazard zones. The environment is depicted in Figure 16. The initial positions of the agent, goal, vases, and hazards are randomized at the beginning of each episode. The reward function is defined as the change in Euclidean distance to the goal between successive steps

$$r_t(s_t, a_t) := d_{t-1} - d_t + \mathbf{1}[d_t \leq \epsilon],$$

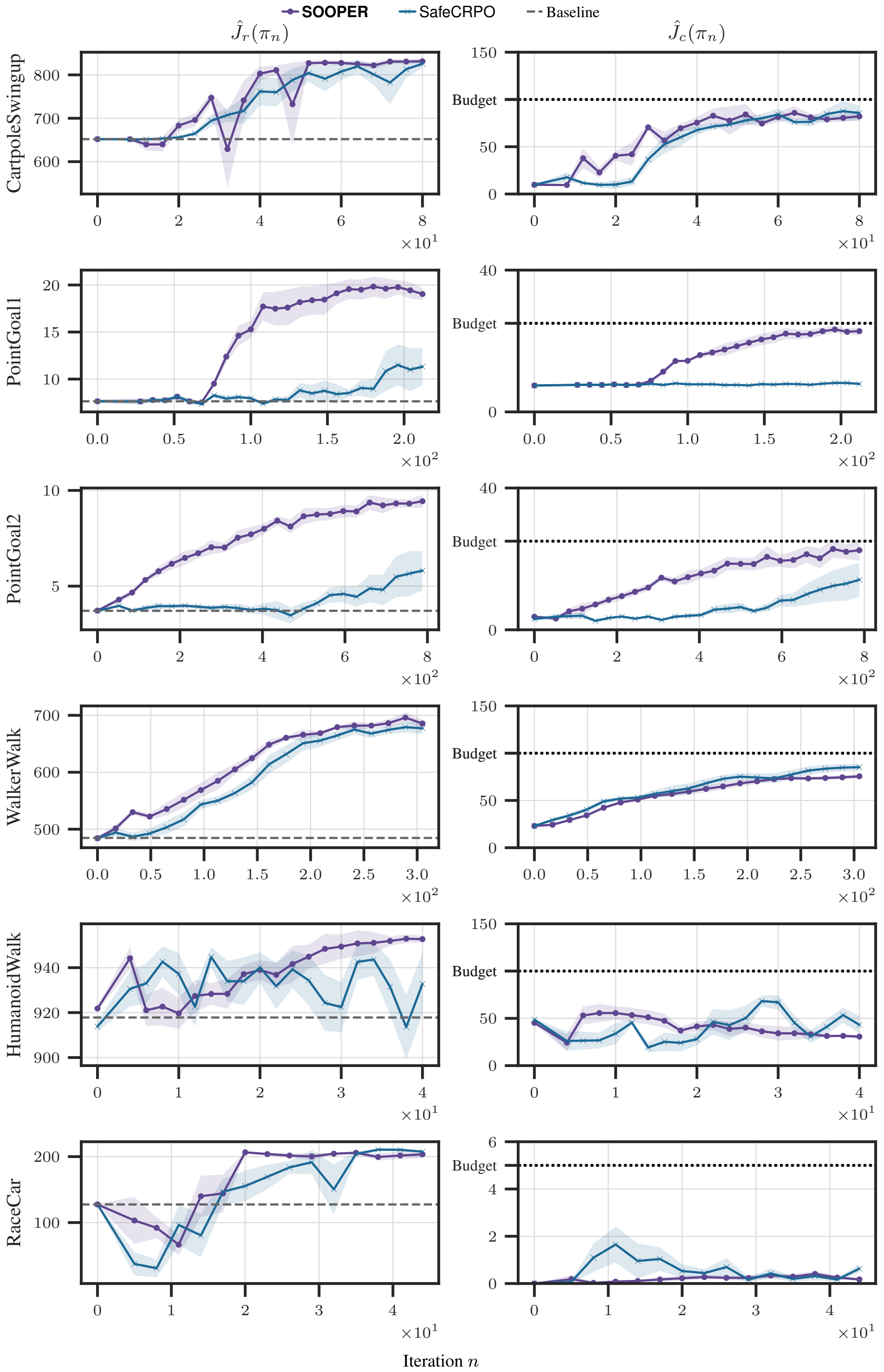


Figure 14: Learning curves of the objective and the constraint for SOOPER and SafeCRPO across all environments. SOOPER consistently outperforms and shows asymptotic convergence.

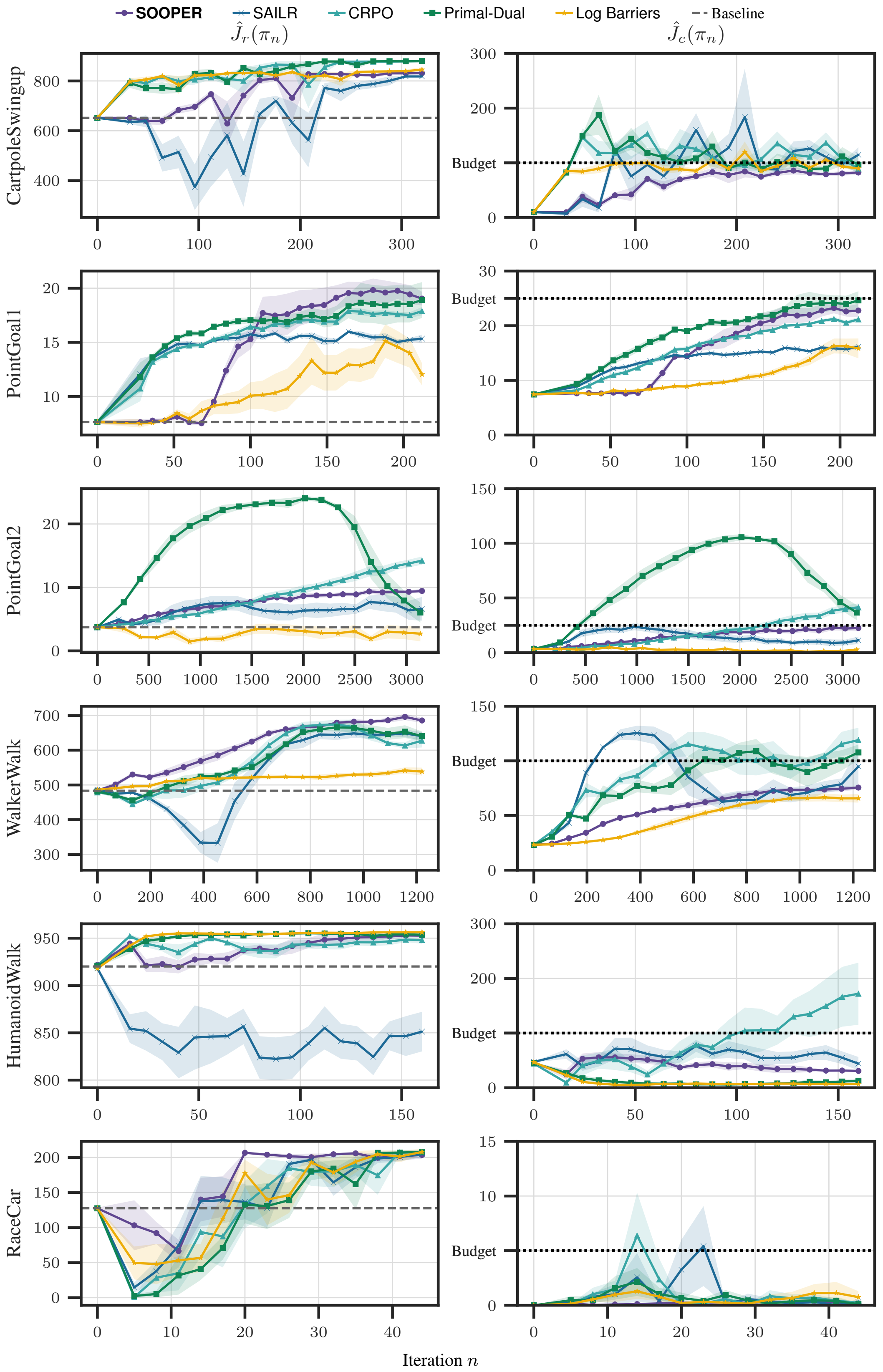


Figure 15: Learning curves of the objective and the constraint for SOOPER and comparison baselines. SOOPER maintains safety throughout training in all tasks while outperforming the baselines in terms of performance.

where $d_t = \|\mathbf{x}_t - \mathbf{x}_{\text{goal}}\|_2$ is the Euclidean distance from the robot to the goal. The term $\mathbf{1}[d_t \leq \epsilon]$ is an indicator that gives a reward bonus when the agent reaches the goal, i.e. when it is within $\epsilon = 0.3$ of the center of the goal. The goal position is resampled to another free position in the environment once reached. PointGoal1 and PointGoal2 differ in the amount of obstacles located in the environment. A cost of 1 is incurred when the agent collides with a vase v, when one of the vases crosses a linear velocity threshold (after collision) or when the agent is inside a hazard zone h:

$$c_t(s_t, a_t) \coloneqq \mathbf{1}[\exists v \in V : \text{collides}(\mathbf{x}_t, \mathbf{x}_v)] + \mathbf{1}[\exists v \in V : \mathbf{x}'_v \geq \gamma] + \mathbf{1}[\exists h \in H : d_t \leq \rho],$$

where $\gamma = 5e^{-2}$ and $\rho = 0.2$. Please see the implementation of Ray et al. [11] for more specific details.

Simulation gap. In our sim-to-sim experiments (in Figure 3), before training, we uniformly sample the gear parameters of the actuators that control the linear and angular velocity of the robot. The precise ranges are given in Table 1. Note that the z-joint is a hinge joint that allows the agent to rotate around the z-axis and the x-joint is a slide joint that allows translation in the xy-plane.

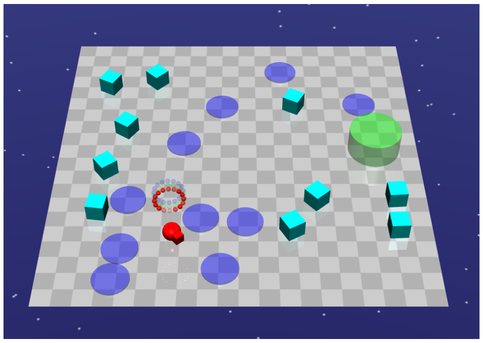


Figure 16: Visualization of a random initialization of the PointGoal2 environment. The red pointmass is the agent, the green transparent cylinder is the goal, the cyan boxes are vases and the blue circles are hazard zones.

Table 1: Domain randomization parameters and ranges used during training.

Parameter	**Value**
Gear (x)	[–0.2, 0.2]
Gear (z)	[–0.1, 0.1]

F RWRL Benchmark

We use the RWRL benchmark suite [10], which adds safety constraints and distribution shifts to the tasks from the DeepMind Control benchmark suite [87].

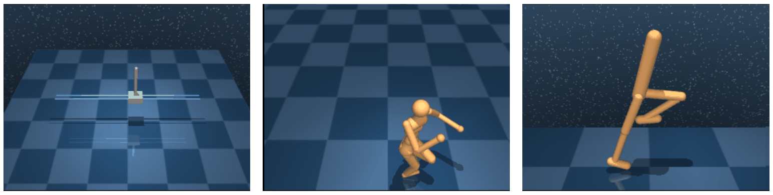


Figure 17: RWRL tasks.

Constraints. We use the joint position limits constraint for HumanoidWalk, joint velocity limits for WalkerWalk, and slider position limits for CartpoleSwingup. These are the standard constraints proposed by Dulac-Arnold et al. [10].

Simulation gap. In our sim-to-sim experiments from Figure 3, we follow a similar experimental setup as Queeney and Benosman [73], introducing variability in the physical properties of the system during training to simulate distribution shifts. In Table 2 we provide the specific parameters we perturb in each task.

G RaceCar Environment

In this environment the agent is asked to navigate to a goal position while avoiding static obstacles. This environment is implemented both in our sim-to-sim experiments, as well as on the real-world race car experiments.

Table 2: Additive domain randomization parameters and ranges used during training of the policy prior across tasks from RWLR.

Parameter	Value
CartpoleSwingup	
Pole Length	[-0.1, 0.1]
Gear	[-1.0, 1.0]
Knee Gear	[-40., 40.0]
HumanoidWalk	
Friction	[-0.05, 0.05]
Hip Gear (x)	[-20.0, 20.0]
Hip Gear (y)	[-20.0, 20.0]
Hip Gear (z)	[-60.0, 60.0]
WalkerWalk	
Torso Length	[-0.1, 0.1]
Gear	[-5.0, 5.0]

Reward and cost. The reward at timestep t is given by

$$r_t(s_t, a_t) := d_{t-1} - d_t + \mathbf{1}[d_t \leq \epsilon] - \lambda_c \|a_t\|_2 - \lambda_l \|a_t - a_{t-1}\|_2^2,$$

where d_t is the Euclidean distance to the goal, and $a_t \in \mathbb{R}^2$ denotes the action applied at time t (consisting of steering and throttle). The term $\mathbf{1}[d_t \leq \epsilon]$ is an indicator function that gives a reward bonus when the agent is within $\epsilon = 0.3$ of the goal. The penalties λ_c and λ_l weight the control effort (magnitude of the action) and the change in action between consecutive timesteps, respectively. The cost function at time t is defined as

$$c_t(s_t, a_t) := \sum_{i=1}^{3} \mathbf{1}\left[\|x_t - p_i\| < \rho_i\right] E_t^k + \mathbf{1}[x_t \notin \mathcal{V}],$$

where $x_t \in \mathbb{R}^2$ is the car's position, p_i and ρ_i are the position and radius of the i-th obstacle and E_t^k is the kinetic energy of the car at time t, simulating a plastic collision between the car and obstacles. This choice of cost function allows us to penalize more severely collisions in which the car smashes with high velocity into obstacles, as opposed to softly touching them. The second term penalizes the agent for leaving the valid area $\mathcal{V}$, which corresponds to a bounded rectangular arena.

Simulation gap. In the previous sim-to-sim environments, we model the sim-to-sim gap for the experiments in Figure 3 by introducing an auxiliary dynamics parameter (e.g., pendulum length) that is not observed during training. In contrast, in the RaceCar environment, the car dynamics in the training environments are governed by a semi-kinematic bicycle model that does not account for interactions between the tire and the ground. On the other hand, in evaluation, we use the dynamical bicycle model of Kabzan et al. [82]. We refer the reader to Kabzan et al. [82] for the detailed equations of motion.